Diffusion maps for changing data[☆]

Ronald R. Coifman, Matthew J. Hirn[*]

Yale University
Department of Mathematics
P.O. Box 208283
New Haven, Connecticut 06520-8283
USA

Abstract

Graph Laplacians and related nonlinear mappings into low dimensional spaces have been shown to be powerful tools for organizing high dimensional data. Here we consider a data set X in which the graph associated with it changes depending on some set of parameters. We analyze this type of data in terms of the diffusion distance and the corresponding diffusion map. As the data changes over the parameter space, the low dimensional embedding changes as well. We give a way to go between these embeddings, and furthermore, map them all into a common space, allowing one to track the evolution of X in its intrinsic geometry. A global diffusion distance is also defined, which gives a measure of the global behavior of the data over the parameter space. Approximation theorems in terms of randomly sampled data are presented, as are potential applications.

Keywords: diffusion distance; graph Laplacian; manifold learning; dynamic graphs; dimensionality reduction; kernel method; spectral graph theory

1. Introduction

In this paper we consider a changing graph depending on certain parameters, such as time, over a fixed set of data points. Given a set of parameters of interest, our goal is to organize the data in such a way that we can perform meaningful comparisons between data points derived from different parameters. In some scenarios, a direct comparison may be possible; on the other hand, the methods we develop are more general and can handle situations in which the changes to the data prevent direct comparisons across the parameter space. In particular, one may consider situations in which the fundamental building blocks of the data set change, perhaps changing the dimension of the data. In order to make meaningful comparisons between different realizations of the data, we look for invariants in the data set as it changes. We model the data set as a

[☆]Submitted to *Applied and Computational Harmonic Analysis* 25 August 2012. LaTeX file last compiled 3 September 2012.

[*]Corresponding author
Email addresses: coifman@math.yale.edu (Ronald R. Coifman), matthew.hirn@yale.edu (Matthew J. Hirn)
URL: www.math.yale.edu/~mh644 (Matthew J. Hirn)

normalized, weighted graph, and measure the similarity between two points based on how the local subgraph around each point changes over the parameter space. The framework we develop will allow for the comparison of any two points derived from any two parameters within the graph, thus allowing one to organize not only along the data points but the parameter space as well.

An example of this type of data comes from hyperspectral image analysis. A hyperspectral image is in fact a set of images of the same scene that are taken at different wavelengths. Put together, these images form a data cube in which the length and width of the cube correspond to spatial dimensions, and the height of the cube corresponds to the different wavelengths. Thus each pixel is in fact a vector corresponding to the spectral signature of the materials contained in that pixel. Consider the situation in which we are given two hyperspectral images of the same scene, and we wish to highlight the anomalous (e.g., man made) changes between the two. Perhaps though, for each data set, different cameras were used which measured different wavelengths, perhaps also at different times of day under different weather conditions. In such a scenario a direct comparison becomes much more difficult. Current work in the field often times goes under the heading change detection, as the goal is to often find small changes in a large scene; see [1] for more details.

Other possible areas for applications come from the modeling of social networks as graphs. The relationships between people change over time and determining how groups of people interact and evolve is a new and interesting problem that has usefulness in marketing and other areas. Financial markets are yet another area that lends itself to analysis conducted over time, as are certain evolutionary biological questions and even medical problems in which patient tests are updated over the course of their lives.

Let $\mathcal{I}$ denote our parameter space, and let X_α, with $\alpha \in \mathcal{I}$, be the data in question. The elements of our data set are fixed, but the graph changes depending on the parameter α. In other words, there is a known bijection between X_α and X_β for $\alpha, \beta \in \mathcal{I}$, but the corresponding graph weights of X have changed between the two parameters. For a fixed α, the diffusion maps framework developed in [2] gives a multiscale way of organizing X_α. More specifically, the diffusion mapping maps X_α into a particular ℓ^2 space in which the usual ℓ^2 distance corresponds to the diffusion distance on X_α. However, for different times α and β, the diffusion map may take X_α and X_β into different ℓ^2 spaces, thus meaning that one cannot take the standard ℓ^2 distance between the elements of these two spaces. Our contribution here is to generalize the diffusion maps framework so that it works independently of the parameter α. In particular, we derive formulas for the distance between points in different embeddings that are in terms of the individual diffusion maps of each space. It is even possible to define a mapping from one embedding to the other, so that after applying this mapping the standard ℓ^2 distance can once again be used to compute diffusion distances. Once this generalized framework has been established, we are able to define a global distance between all of X_α and X_β based on the behavior of the diffusions within each data set. This distance in turn allows one to model the global behavior of X_α as it changes over $\mathcal{I}$.

Earlier results that use diffusion maps to compare two data sets can be found in [3]. Furthermore, there is recent work contained in [4] That also involves combining diffusion geometry principles via tree structures with evolving graphs. In [5], the author considers the case of an evolving Riemannian manifold on which a diffusion process is spreading as the manifold evolves. In our work, we separate out the two processes, effectively using the diffusion process to organize the evolution of the data. Also tangentially related to this work are the results contained in [6] on shape analysis, in which shapes are compared via their heat kernels. More generally,

this paper fits into the larger class of research that utilizes nonlinear mappings into low dimensional spaces in order to organize potentially high dimensional data; examples include locally linear embedding (LLE) [7], ISOMAP [8], Hessian LLE [9], Laplacian eigenmaps [10], and the aforementioned diffusion maps [2].

An outline of this paper goes as follows: in the next section, we take care of some notation and review the diffusion mapping first presented in [2]. In Section 3 we generalize the diffusion distance for a data set that changes over some parameter space, and show that it too can be computed in terms the spectral embeddings of the corresponding diffusion operators. We also show how to map each of the embeddings into one common embedding in which the ℓ^2 distance is equal to the diffusion distance. The global diffusion distance between graphs is defined in Section 4; it is also seen to be able to be computed in terms of the eigenvalues and eigenfunctions of the relevant diffusion operators. In Section 5 we set up and state two random sampling theorems, one for the diffusion distance and one for the global diffusion distance. The proofs of these theorems are given in Appendix B. Section 6 contains some applications, and we conclude with some remarks and possible future directions in Section 7.

2. Notation and preliminaries

In this section we introduce some basic notation and review certain preliminary results that will motivate our work.

2.1. Notation

Let $\mathbb{R}$ denote the real numbers and let $\mathbb{N} \triangleq \{1, 2, 3, \ldots\}$. Often we will use constants that depend on certain variables or parameters. We let $C(\cdot)$, $C_1(\cdot)$, $C_2(\cdot)$, etc, denote these constants; note that they can change from line to line.

We recall some basic notation from operator theory. Let $\mathcal{H}$ be a real, separable Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. Let $A : \mathcal{H} \to \mathcal{H}$ be a bounded, linear operator, and let A^* be its adjoint. The operator norm of A is defined as:

$$\|A\| \triangleq \sup_{\|f\|=1} \|Af\|.$$

A bounded operator A is Hilbert-Schmidt if

$$\sum_{i \geq 1} \|Ae^{(i)}\|^2 < \infty$$

for some (and hence any) Hilbert basis $\{e^{(i)}\}_{i \geq 1}$. The space of Hilbert-Schmidt operators is also a Hilbert space with scalar product

$$\langle A, B \rangle_{HS} \triangleq \sum_{i \geq 1} \langle Ae^{(i)}, Be^{(i)} \rangle.$$

We denote the corresponding norm as $\| \cdot \|_{HS}$. Note that if an operator is Hilbert-Schmidt, then it is compact. A subset of the Hilbert-Schmidt operators are those that are trace class. A bounded operator A is trace class if

$$\sum_{i \geq 1} \langle \sqrt{A^* A} e^{(i)}, e^{(i)} \rangle < \infty$$

for some (and hence any) Hilbert basis $\{e^{(i)}\}_{i\geq 1}$. For any trace class operator A, we have

$$\mathrm{Tr}(A) \triangleq \sum_{i\geq 1} \langle Ae^{(i)}, e^{(i)}\rangle < \infty,$$

where $\mathrm{Tr}(A)$ is called the trace of A. The space of trace class operators is a Banach space endowed with the norm

$$\|A\|_{TC} \triangleq \mathrm{Tr}(\sqrt{A^*A}).$$

Note that the different operator norms are related as follows:

$$\|A\| \leq \|A\|_{HS} \leq \|A\|_{TC}.$$

2.2. Diffusion maps

In this section we consider just a single data set that does not change with some parameter and review the notion of diffusion maps on this data set. We assume that we are given a measure space (X, μ), consisting of data points X that are distributed according to μ. The diffusion maps framework developed in [2] gives a multiscale organization of the data set X. We also have a positive, symmetric kernel $k : X \times X \to \mathbb{R}$ that encodes how similar two data points are. From X and k, one can construct a weighted graph $\Gamma \triangleq (X, k)$, in which the vertices of Γ are the data points $x \in X$, and the weight of the edge xy is given by $k(x, y)$.

Define the density, $m : X \to \mathbb{R}$, as

$$m(x) \triangleq \int_X k(x, y)\, d\mu(y), \quad \text{for all } x \in X. \tag{1}$$

We assume that the density m satisfies

$$m(x) > 0, \quad \text{for } \mu \text{ a.e. } x \in X, \tag{2}$$

and

$$m \in L^1(X, \mu). \tag{3}$$

Given (2), the weight function

$$p(x, y) \triangleq \frac{k(x, y)}{m(x)}$$

is well defined for $\mu \otimes \mu$-a.e. $(x, y) \in X \times X$. Although p is no longer symmetric, it does satisfy the following useful property:

$$\int_X p(x, y)\, d\mu(y) = 1, \quad \text{for } \mu \text{ a.e. } x \in X.$$

Therefore we can view p as the transition kernel of a Markov chain on X. Equivalently, if $p \in L^2(X \times X, \mu \otimes \mu)$, the integral operator $P : L^2(X, \mu) \to L^2(X, \mu)$, defined as

$$(Pf)(x) \triangleq \int_X p(x, y) f(y)\, d\mu(y), \quad \text{for all } f \in L^2(X, \mu),$$

is a diffusion operator. In particular, the value $p(x, y)$ represents the probability of transition in one time step from the vertex x to the vertex y, which is proportional to the edge weight $k(x, y)$.

For $t \in \mathbb{N}$, let $p^{(t)}(x, y)$ represent the probability of transition in t time steps from the node x to the node y; note that $p^{(t)}$ is the kernel of the operator P^t. As shown in [2], running the Markov chain forward, or equivalently taking powers of P, reveals relevant geometric structures of X at different scales. In particular, small powers of P will segment the data set into several smaller clusters. As t is increased and the Markov chain diffuses across the graph Γ, the clusters evolve and merge together until in the limit as $t \to \infty$ the data set is grouped into one cluster (assuming the graph is connected).

The phenomenon described above can be encapsulated by the diffusion distance at time t between two vertices x and y in the graph Γ. In order to define the diffusion distance, we first note that the Markov chain constructed above has the stationary distribution $\pi : X \to \mathbb{R}$, where

$$\pi(x) = \frac{m(x)}{\int_X m(y)\, d\mu(y)}.$$

Combining (2) and (3) we see that $\pi(x)$ is well defined for μ a.e. $x \in X$. The diffusion distance between $x, y \in X$ is then defined as:

$$\begin{aligned} \widetilde{D}^{(t)}(x, y)^2 &\triangleq \left\| p^{(t)}(x, \cdot) - p^{(t)}(y, \cdot) \right\|^2_{L^2(X, d\mu/\pi)} \\ &= \int_X \left(p^{(t)}(x, u) - p^{(t)}(y, u) \right)^2 \frac{d\mu(u)}{\pi(u)}. \end{aligned}$$

A simplified formula for the diffusion distance can be found by considering the spectral decomposition of P. Define the kernel $a : X \times X \to \mathbb{R}$ as

$$a(x, y) \triangleq \frac{\sqrt{m(x)}}{\sqrt{m(y)}} p(x, y) = \frac{k(x, y)}{\sqrt{m(x)}\sqrt{m(y)}}, \quad \text{for } \mu \otimes \mu \text{ a.e. } (x, y) \in X \times X.$$

If $a \in L^2(X \times X, \mu \otimes \mu)$, then P has a discrete set of eigenfunctions $\{\upsilon^{(i)}\}_{i \geq 1}$ with corresponding eigenvalues $\{\lambda^{(i)}\}_{i \geq 1}$. It can then be shown that

$$\widetilde{D}^{(t)}(x, y)^2 = \sum_{i \geq 1} \left(\lambda^{(i)} \right)^{2t} \left(\upsilon^{(i)}(x) - \upsilon^{(i)}(y) \right)^2. \tag{4}$$

Inspired by (4), [2] defines the diffusion map $\Upsilon^{(t)} : X \to \ell^2$ at diffusion time t to be:

$$\Upsilon^{(t)}(x) \triangleq \left(\left(\lambda^{(i)} \right)^t \upsilon^{(i)}(x) \right)_{i \geq 1}.$$

Therefore, the diffusion distance at time t between $x, y \in X$ is equal to the ℓ^2 norm of the difference between $\Upsilon^{(t)}(x)$ and $\Upsilon^{(t)}(y)$:

$$\widetilde{D}^{(t)}(x, y) = \left\| \Upsilon^{(t)}(x) - \Upsilon^{(t)}(y) \right\|_{\ell^2}.$$

One can also define a second diffusion distance in terms of the symmetric kernel a as opposed to the asymmetric kernel p. In particular, define the operator $A : L^2(X, \mu) \to L^2(X, \mu)$ as

$$(Af)(x) \triangleq \int_X a(x, y) f(y)\, d\mu(y), \quad \text{for all } f \in L^2(X, \mu).$$

Like the diffusion operator P, the operator A and its powers, A^t, reveal the relevant geometric structures of the data set X. Letting $a^{(t)} : X \times X \to \mathbb{R}$ denote the kernel of the operator A^t, we can define another diffusion distance $D^{(t)} : X \times X \to \mathbb{R}$ as follows:

$$\begin{aligned} D^{(t)}(x,y)^2 &\triangleq \left\| a^{(t)}(x,\cdot) - a^{(t)}(y,\cdot) \right\|^2_{L^2(X,\mu)} \\ &= \int_X \left(a^{(t)}(x,u) - a^{(t)}(y,u) \right)^2 d\mu(u). \end{aligned}$$

As before, we consider the spectral decomposition of A. Let $\{\lambda^{(i)}\}_{i\geq 1}$ and $\{\psi^{(i)}\}_{i\geq 1}$ denote the eigenvalues and eigenfunctions of A (indeed, the nonzero eigenvalues of P and A are the same), and define the diffusion map $\Psi^{(t)} : X \to \ell^2$ (corresponding to A) as

$$\Psi^{(t)}(x) = \left(\left(\lambda^{(i)}\right)^t \psi^{(i)}(x) \right)_{i\geq 1}.$$

Then, under the same assumptions as before, we have

$$D^{(t)}(x,y)^2 = \left\| \Psi^{(t)}(x) - \Psi^{(t)}(y) \right\|^2_{\ell^2} = \sum_{i\geq 1} \left(\lambda^{(i)}\right)^t \left(\psi^{(i)}(x) - \psi^{(i)}(y) \right)^2. \tag{5}$$

We make a few remarks concerning the differences between the two formulations. First, we note that the original diffusion distance $\widetilde{D}^{(t)}$ is defined as an L^2 distance under the weighted measure $d\mu/\pi$. The second diffusion distance, $D^{(t)}$, due to the symmetric normalization built into the kernel a, is defined only in terms of the underlying measure μ. Furthermore, the eigenfunctions of A are orthogonal, unlike the eigenfunctions of P. Finally, as we have already noted, the eigenvalues of P and A are in fact the same, and furthermore they are contained in $(-1, 1]$. If the graph Γ is connected, then the eigenfunction of P with eigenvalue one is simply the function that maps every element of X to one. The corresponding eigenfunction of A though is the square root of the density, i.e., $\sqrt{m(x)}$. Thus, while both versions of the diffusion distance merge smaller clusters into large clusters as t grows, $\widetilde{D}^{(t)}$ will merge every data point into the same cluster in the limit as $t \to \infty$, while $D^{(t)}$ will reflect the behavior of the density m in the limit as $t \to \infty$.

3. Generalizing the diffusion distance for changing data

In this section we generalize the diffusion maps framework for data sets with input parameters.

3.1. Defining the diffusion distance on a family of graphs

We now turn our attention to the original problem introduced at the beginning of this paper. In its most general form, we are given a parameter space $\mathcal{I}$ and a data set X_α that depends on $\alpha \in \mathcal{I}$. The data points of X_α are given by x_α. The parameter space $\mathcal{I}$ can be continuous, discrete, or completely arbitrary. Since there is an obvious and a priori known bijective correspondence between X_α and X_β for any $\alpha, \beta \in \mathcal{I}$, we consider the following model throughout the remainder of this paper. We are given a single measure space (X, μ) that we think of as changing over $\mathcal{I}$. The measure μ here represents some underlying distribution of the points in X that does not change over $\mathcal{I}$. The evolution of (X, μ) is completely encoded by a family of kernels $k_\alpha : X \times X \to \mathbb{R}$, which measure the similarity between two data points $x, y \in X$ for the parameter $\alpha \in \mathcal{I}$.

Our goal is to reveal the relevant geometric structures of X across the entire parameter space $\mathcal{I}$, and to furthermore have a way of comparing structures from one parameter to other structures derived from a second parameter. To do so, we shall generalize the diffusion distance so that we can compare diffusions derived from different parameters. The first step is to once again consider each pairing X and k_α as a weighted graph, which we denote as $\Gamma_\alpha \triangleq (X, k_\alpha)$.

Updating our notation for this dynamic setting, for each parameter $\alpha \in \mathcal{I}$ we have the density $m_\alpha : X \to \mathbb{R}$ defined as

$$m_\alpha(x) \triangleq \int_X k_\alpha(x, y)\, d\mu(y), \quad \text{for all } \alpha \in \mathcal{I}, \;\; x \in X.$$

For reasons that shall become clear later, we slightly strengthen the assumptions on m_α as compared to those in equations (2) and (3). In particular, we assume that

$$m_\alpha(x) > 0, \quad \text{for all } \alpha \in \mathcal{I}, \;\; x \in X,$$

and

$$m_\alpha \in L^1(X, \mu), \quad \text{for all } \alpha \in \mathcal{I}.$$

We then define two classes of kernels $a_\alpha : X \times X \to \mathbb{R}$ and $p_\alpha : X \times X \to \mathbb{R}$ in the same manner as earlier:

$$a_\alpha(x, y) \triangleq \frac{k_\alpha(x, y)}{\sqrt{m_\alpha(x)}\sqrt{m_\alpha(y)}}, \quad \text{for all } \alpha \in \mathcal{I}, \;\; (x, y) \in X \times X, \tag{6}$$

and

$$p_\alpha(x, y) \triangleq \frac{k_\alpha(x, y)}{m_\alpha(x)}, \quad \text{for all } \alpha \in \mathcal{I}, \;\; (x, y) \in X \times X.$$

Assume that $a_\alpha, p_\alpha \in L^2(X \times X, \mu \otimes \mu)$. Their corresponding integral operators are given by $A_\alpha : L^2(X, \mu) \to L^2(X, \mu)$ and $P_\alpha : L^2(X, \mu) \to L^2(X, \mu)$, where

$$(A_\alpha f)(x) \triangleq \int_X a_\alpha(x, y) f(y)\, d\mu(y), \quad \text{for all } \alpha \in \mathcal{I}, \;\; f \in L^2(X, \mu), \tag{7}$$

and

$$(P_\alpha f)(x) \triangleq \int_X p_\alpha(x, y) f(y)\, d\mu(y), \quad \text{for all } \alpha \in \mathcal{I}, \;\; f \in L^2(X, \mu).$$

Finally, we let $a_\alpha^{(t)}$ and $p_\alpha^{(t)}$ denote the kernels of the integral operators A_α^t and P_α^t, respectively.

Returning to the task at hand, in order to compare Γ_α with Γ_β, it is possible to use the operators A_α and A_β or P_α and P_β. We choose to perform our analysis using the symmetric operators, as it shall simplify certain things. For now, consider the function $a_\alpha(x, \cdot)$ for a fixed $x \in X$. We think of this function in the following way. Consider the graph Γ_α, and imagine dropping a unit of mass on the node x and allowing it to spread, or diffuse, throughout Γ_α. After one unit of time, the amount of mass that has spread from x to some other node y is proportional to $a_\alpha(x, y)$. Similarly, if we want to let the mass spread throughout the graph for a longer period of time, we can, and the amount of mass that has spread from x to y after t units of time is then proportional to $a_\alpha^{(t)}(x, y)$. The diffusion distance at time t, which is the L^2 norm of $a_\alpha^{(t)}(x, \cdot) - a_\alpha^{(t)}(y, \cdot)$, is then comparing the behavior of the diffusion centered at x with the behavior of the diffusion centered

at y. We wish to extend this idea for different parameters α and β. In other words, we wish to have a meaningful distance between x at parameter α and y at parameter β that is based on the same principle of measuring how their respective diffusions behave.

Our solution is to generalize the diffusion distance in the following way. For each diffusion time $t \in \mathbb{N}$, we define a dynamic diffusion distance $D^{(t)} : (X \times \mathcal{I}) \times (X \times \mathcal{I}) \to \mathbb{R}$ as follows. Let $x_\alpha \triangleq (x, \alpha) \in X \times \mathcal{I}$, and set

$$\begin{aligned} D^{(t)}(x_\alpha, y_\beta)^2 &\triangleq \left\| a_\alpha^{(t)}(x, \cdot) - a_\beta^{(t)}(y, \cdot) \right\|_{L^2(X,\mu)}^2 \\ &= \int_X \left(a_\alpha^{(t)}(x, u) - a_\beta^{(t)}(y, u) \right)^2 d\mu(u). \end{aligned}$$

This notion of distance can be thought of as comparing how the neighborhood of x_α differs from the neighborhood of y_β. In particular, if we are comparing the same data point but at different parameters, for example x_α and x_β, the diffusion distance between them will be small if their neighborhoods do not change much from α to β. On the other hand, if say a large change occurs at x at parameter β, then the neighborhood of x_β should differ from the neighborhood of x_α and so they will have a large diffusion distance between them.

Some more intuition about the quantity $D^{(t)}(x_\alpha, y_\beta)$ can be derived from the triangle inequality. In particular, one application of it gives

$$D^{(t)}(x_\alpha, y_\beta) \leq D^{(t)}(x_\alpha, x_\beta) + D^{(t)}(x_\beta, y_\beta).$$

Thus we see that $D^{(t)}(x_\alpha, y_\beta)$ is bounded from above by the change in x from α to β (i.e. the quantity $D^{(t)}(x_\alpha, x_\beta)$) plus the diffusion distance between x and y in the graph Γ_β (i.e. the quantity $D^{(t)}(x_\beta, y_\beta)$).

Remark 3.1. As noted earlier, we have chosen to generalize the diffusion distance in terms of the symmetric kernels a_α as opposed to the asymmetric kernels p_α. The primary reason for this choice is that when using the kernel p_α to compute the diffusion distance between x and y, we must use the weighted measure $d\mu/\pi_\alpha$, where π_α denotes the stationary distribution of the Markov chain on Γ_α. Thus, when computing the diffusion distance between x_α and y_β, one must incorporate this weighted measure as well. Since the stationary distribution will invariably change from α to β, the most natural generalization in this case would be:

$$\widetilde{D}^{(t)}(x_\alpha, y_\beta)^2 \triangleq \int_X \left(\frac{p_\alpha^{(t)}(x, u)}{\sqrt{\pi_\alpha(u)}} - \frac{p_\beta^{(t)}(y, u)}{\sqrt{\pi_\beta(u)}} \right)^2 d\mu(u).$$

3.2. Diffusion maps for $\{\Gamma_\alpha\}_{\alpha \in \mathcal{I}}$

Analogous to the diffusion distance for a single graph $\Gamma = (X, k)$, we can write the diffusion distance for $\{\Gamma_\alpha\}_{\alpha \in \mathcal{I}}$ in terms the spectral decompositions of $\{A_\alpha\}_{\alpha \in \mathcal{I}}$. We first collect the following mild, but necessary, assumptions, some of which have already been stated.

Assumption 1. *We assume the following properties:*

1. *$X \subseteq \mathbb{R}^d$ and μ is a σ-finite measure.*
2. *The kernel k_α is positive definite and symmetric for all $\alpha \in \mathcal{I}$.*

3. *For each $\alpha \in \mathcal{I}$, $m_\alpha \in L^1(X,\mu)$ and $m_\alpha > 0$.*
4. *For any $\alpha \in \mathcal{I}$, the operator A_α is trace class.*

A few remarks concerning the assumed properties. First, the reader may have noticed that we replaced the assumption that k_α be positive with the stronger assumption that it is positive definite. This combined with the third property that $m_\alpha(x) > 0$ for all $x \in X$, implies that a_α is also positive definite. Thus the operators A_α are positive and self adjoint.

If one wished to revert back to the weaker assumption that k_α merely be positive, then the following adjustment could be made. Clearly the symmetrically normalized kernel a_α will still be positive, but the operator A_α may not be. However, one could replace A_α, for each $\alpha \in \mathcal{I}$, with the graph Laplacian $L_\alpha : L^2(X,\mu) \to L^2(X,\mu)$, which is defined as

$$L_\alpha \triangleq \frac{1}{2}(I - A_\alpha),$$

where $I : L^2(X,\mu) \to L^2(X,\mu)$ is the identity operator. The graph Laplacian L_α is a positive operator with eigenvalues contained in $[0,1]$. The analysis that follows would still apply with only minor adjustments.

The fourth item that A_α be trace class plays a key role in the results of this section, and itself implies that these operators are Hilbert-Schmidt and so also compact. Thus, as a further consequence, $a_\alpha \in L^2(X \times X, \mu \otimes \mu)$ for each $\alpha \in \mathcal{I}$.

We note that assumptions three and four are both satisfied if for each $\alpha \in \mathcal{I}$ the kernel k_α is continuous, bounded from above and below, and if the measure of X is finite. That is, if for each α,

$$0 < C_1(\alpha) \leq k_\alpha(x,y) \leq C_2(\alpha) < \infty, \quad \text{for all } (x,y) \in X \times X,$$

and

$$\mu(X) < \infty,$$

then we can derive assumptions three and four.

As an immediate consequence of the properties contained in Assumption 1, we see that for each α the operator A_α has a countable collection of positive eigenvalues and orthonormal eigenfunctions. Let $\{\lambda_\alpha^{(i)}\}_{i\geq 1}$ and $\{\psi_\alpha^{(i)}\}_{i\geq 1}$ be the eigenvalues and a set of orthonormal eigenfunctions of A_α, respectively, so that

$$(A_\alpha \psi_\alpha^{(i)})(x) = \lambda_\alpha^{(i)} \psi_\alpha^{(i)}(x), \quad \text{for } \mu \text{ a.e. } x \in X,$$

and

$$\langle \psi_\alpha^{(i)}, \psi_\alpha^{(j)} \rangle_{L^2(X,\mu)} = \delta(i-j), \quad \text{for all } i,j \geq 1.$$

Furthermore, as noted in [2], the eigenvalues of P_α are bounded in absolute value by one, with at least one eigenvalue equaling one. Since the eigenvalues of A_α and P_α are the same, we also have

$$1 = \lambda_\alpha^{(1)} \geq \lambda_\alpha^{(2)} \geq \lambda_\alpha^{(3)} \geq \ldots,$$

where $\lambda_\alpha^{(i)} \to 0$ as $i \to \infty$.

As with the original diffusion distance defined on a single data set, our generalized notion of the diffusion distance for dynamic data sets has a simplified form in terms of the spectral decompositions of the relevant operators.

Theorem 3.2. *Let (X, μ) be a measure space and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ a family of kernels defined on X. If (X, μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the properties of Assumption 1, then the diffusion distance at time t between x_α and y_β can be written as:*

$$D^{(t)}(x_\alpha, y_\beta)^2 = \sum_{i\geq 1} \left(\lambda_\alpha^{(i)}\right)^{2t} \psi_\alpha^{(i)}(x)^2 + \sum_{j\geq 1} \left(\lambda_\beta^{(j)}\right)^{2t} \psi_\beta^{(j)}(y)^2 - 2\sum_{i,j\geq 1} \left(\lambda_\alpha^{(i)}\right)^{t} \left(\lambda_\beta^{(j)}\right)^{t} \psi_\alpha^{(i)}(x)\, \psi_\beta^{(j)}(y)\, \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)}, \tag{8}$$

where for each pair $(\alpha, \beta) \in \mathcal{I} \times \mathcal{I}$, equation (8) *converges in $L^2(X \times X, \mu \otimes \mu)$. If, additionally, k_α is continuous for each $\alpha \in \mathcal{I}$, $X \subseteq \mathbb{R}^d$ is closed, and μ is a strictly positive Borel measure, then* (8) *holds for all $(x, y) \in X \times X$.*

We postpone the proof of Theorem 3.2 till the end of this section. Notice that equation (8) is in fact an extension of the formula given for the diffusion distance on a single data set. Indeed, if one were to take x_α and $y_\beta = y_\alpha$, the formula given in (8) would simplify to (5) with the underlying kernel taken to be k_α. Thus, it is natural to define the diffusion map $\Psi_\alpha^{(t)} : X \to \ell^2$ for the parameter α and diffusion time t as

$$\Psi_\alpha^{(t)}(x) \triangleq \left(\left(\lambda_\alpha^{(i)}\right)^t \psi_\alpha^{(i)}(x) \right)_{i\geq 1}. \tag{9}$$

For $v \in \ell^2$, let $v[i]$ denote the i^{th} element of the sequence u. Using (9), one can write equation (8) as

$$D^{(t)}(x_\alpha, y_\beta)^2 = \left\|\Psi_\alpha^{(t)}(x)\right\|_{\ell^2}^2 + \left\|\Psi_\beta^{(t)}(y)\right\|_{\ell^2}^2 - 2\sum_{i,j\geq 1} \Psi_\alpha^{(t)}(x)[i]\, \Psi_\beta^{(t)}(y)[j]\, \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)}. \tag{10}$$

In particular, one has in general that

$$D^{(t)}(x_\alpha, y_\beta) \neq \left\|\Psi_\alpha^{(t)}(x) - \Psi_\beta^{(t)}(y)\right\|_{\ell^2}.$$

Intuitively, the thing to take away from this discussion is that for each parameter $\alpha \in \mathcal{I}$, the diffusion map $\Psi_\alpha^{(t)}$ maps X into an ℓ^2 space that itself also depends on α. The ℓ^2 embedding corresponding to α is not the same as the ℓ^2 embedding corresponding to $\beta \in \mathcal{I}$, but equation (10) gives a way of computing distances between the different ℓ^2 embeddings.

Also, once again paralleling the original diffusion distance, we see that if the eigenvalues of A_α and A_β decay sufficiently fast, then the diffusion distance can be well approximated by a small, finite number of eigenvalues and eigenfunctions of these two operators. In particular, we need only map Γ_α and Γ_β into finite dimensional Euclidean spaces.

Remark 3.3. One interesting aspect of the diffusion distance is its asymptotic behavior as $t \to \infty$, and in particular that behavior when the family of graphs $\{\Gamma_\alpha\}_{\alpha\in\mathcal{I}}$ are connected graphs. In this case, each operator A_α has precisely one eigenvalue equal to one, and the corresponding eigenfunction is the square root of the density, i.e.,

$$1 = \lambda_\alpha^{(1)} > \lambda_\alpha^{(2)} \geq \lambda_\alpha^{(3)} \geq \ldots, \quad \text{and} \quad \psi_\alpha^{(1)} = \sqrt{m_\alpha(\cdot)}.$$

It is quite simple to show then that

$$\lim_{t\to\infty} D^{(t)}(x_\alpha, y_\beta)^2 = \left(\sqrt{m_\alpha(x)} - \sqrt{m_\beta(y)} \right)^2 + \sqrt{m_\alpha(x)}\sqrt{m_\beta(y)} \left\| \sqrt{m_\alpha(\cdot)} - \sqrt{m_\beta(\cdot)} \right\|_{L^2(X,\mu)}^2.$$

Thus the asymptotic diffusion distance can be computed without diagonalizing any of the diffusion operators. Furthermore, we see that it is not just the pointwise difference between the densities, but rather it is the pointwise difference plus a term that takes into account the global difference between the two densities. It can be used as a fast way of determing significant changes from α to β.

Proof of Theorem 3.2. We first use the fact that for each $\alpha \in \mathcal{I}$, A_α is a positive, trace class operator. Thus A_α is Hilbert-Schmidt, and so we know that for each $\alpha \in \mathcal{I}$,

$$a_\alpha^{(t)}(x,y) = \sum_{i\geq 1} \left(\lambda_\alpha^{(i)}\right)^t \psi_\alpha^{(i)}(x)\,\psi_\alpha^{(i)}(y), \quad \text{with convergence in } L^2(X\times X, \mu\otimes\mu). \tag{11}$$

If the additional assumptions hold that k_α is continuous, X is a closed subset of $\mathbb{R}^d$, and μ is a strictly positive Borel measure, then by Mercer's Theorem (see [11, 12]) equation (11) will hold for all $(x,y) \in X \times X$. In this case the proof can be easily amended to get the stronger result; we omit the details.

Expand the formula for $D^{(t)}(x_\alpha, y_\beta)$ as follows:

$$D^{(t)}(x_\alpha, y_\beta)^2 = \int_X \left(a_\alpha^{(t)}(x,u)^2 - 2a_\alpha^{(t)}(x,u)\,a_\beta^{(t)}(y,u) + a_\beta^{(t)}(y,u)^2\right) d\mu(u). \tag{12}$$

We shall evaluate each of the three terms in (12) separately. For the cross term we have,

$$\int_X a_\alpha^{(t)}(x,u)\,a_\beta^{(t)}(y,u)\,d\mu(u) = \int_X \left(\sum_{i,j\geq 1} \left(\lambda_\alpha^{(i)}\right)^t \left(\lambda_\beta^{(j)}\right)^t \psi_\alpha^{(i)}(x)\,\psi_\beta^{(j)}(y)\,\psi_\alpha^{(i)}(u)\,\psi_\beta^{(j)}(u)\right) d\mu(u), \tag{13}$$

with convergence in $L^2(X\times X, \mu\otimes\mu)$. At this point we would like to switch the integral and the summation in line (13); this can be done by applying Theorem 2.25, page 55, from [13], which requires one to show the following:

$$\sum_{i,j\geq 1} \int_X \left|\left(\lambda_\alpha^{(i)}\right)^t \left(\lambda_\beta^{(j)}\right)^t \psi_\alpha^{(i)}(x)\,\psi_\beta^{(j)}(y)\,\psi_\alpha^{(i)}(u)\,\psi_\beta^{(j)}(u)\right| d\mu(u) < \infty. \tag{14}$$

One can prove (14) for $\mu\otimes\mu$ almost every $(x,y) \in X\times X$ through the use of Hölder's Theorem and the fact that we assumed that A_α is a trace class operator for each $\alpha \in \mathcal{I}$; we leave the details to the reader. Thus for $\mu\otimes\mu$ almost every $(x,y) \in X\times X$ we can switch the integral and the summation in line (13), which gives:

$$\int_X a_\alpha^{(t)}(x,u)\,a_\beta^{(t)}(y,u)\,d\mu(u) = \sum_{i,j\geq 1} \left(\lambda_\alpha^{(i)}\right)^t \left(\lambda_\beta^{(j)}\right)^t \psi_\alpha^{(i)}(x)\,\psi_\beta^{(j)}(y)\,\langle \psi_\alpha^{(i)}, \psi_\beta^{(j)}\rangle_{L^2(X,\mu)}, \tag{15}$$

again with convergence in $L^2(X\times X, \mu\otimes\mu)$. A similar calculation shows that, for each $\alpha \in \mathcal{I}$,

$$\int_X a_\alpha^{(t)}(x,u)^2\,d\mu(u) = \sum_{i\geq 1} \left(\lambda_\alpha^{(i)}\right)^{2t} \psi_\alpha^{(i)}(x)^2, \quad \text{with convergence in } L^2(X,\mu). \tag{16}$$

Combining equations (15) and (16) we arrive at the desired formula for $D^{(t)}(x_\alpha, y_\beta)$. □

3.3. Mapping one diffusion embedding into another

As mentioned in the previous subsection, the diffusion map $\Psi_\alpha^{(t)}$ takes X into an ℓ^2 space that itself depends on α. While (10) gives a way of computing distances between two diffusion embeddings, it is also possible to map the embedding $\Psi_\beta^{(t)}(X)$ into the ℓ^2 space of $\Psi_\alpha^{(t)}(X)$. Furthermore, the operator that does so is quite simple. The eigenfunctions $\{\psi_\alpha^{(i)}\}_{i\geq 1}$ are essentially a basis for the embedding of X with parameter α, while the eigenfunctions $\{\psi_i^{(\beta)}\}_{i\geq 1}$ are essentially a basis for the embedding of X with parameter β. The operator that maps one space into the other is similar to the change of basis operator. Define $O_{\beta\to\alpha} : \ell^2 \to \ell^2$ as

$$O_{\beta\to\alpha} v \triangleq \left(\sum_{j\geq 1} v[j] \, \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)} \right)_{i\geq 1}, \quad \text{for all } v \in \ell^2.$$

By the Spectral Theorem, we know that the eigenfunctions of A_α can be taken to form an orthonormal basis for $L^2(X,\mu)$. Thus, the operator $O_{\alpha\to\beta}$ preserves inner products. Indeed, define the operator $S_\alpha : L^2(X,\mu) \to \ell^2$ as

$$S_\alpha f \triangleq \left(\langle \psi_\alpha^{(i)}, f \rangle_{L^2(X,\mu)} \right)_{i\geq 1}, \quad \text{for all } f \in L^2(X,\mu).$$

The adjoint of S_α, $S_\alpha^* : \ell^2 \to L^2(X,\mu)$, is then given by

$$S_\alpha^* v = \sum_{i\geq 1} v[i] \, \psi_\alpha^{(i)}, \quad \text{for all } v \in \ell^2.$$

Since $\{\psi_\alpha^{(i)}\}_{i\geq 1}$ is an orthonormal basis for $L^2(X,\mu)$, $S_\alpha^* S_\alpha = I_{L^2(X,\mu)}$. Therefore, for any $v, w \in \ell^2$,

$$\begin{aligned}
\langle O_{\beta\to\alpha} v, O_{\beta\to\alpha} w \rangle_{\ell^2} &= \sum_{j,k\geq 1} v[j] \, w[k] \left(\sum_{i\geq 1} \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)} \langle \psi_\alpha^{(i)}, \psi_\beta^{(k)} \rangle_{L^2(X,\mu)} \right) \\
&= \sum_{j,k\geq 1} v[j] \, w[k] \, \langle S_\alpha \psi_\beta^{(j)}, S_\alpha \psi_\beta^{(k)} \rangle_{\ell^2} \\
&= \sum_{j,k\geq 1} v[j] \, w[k] \, \delta(j-k) \\
&= \langle v, w \rangle_{\ell^2} \qquad (17)
\end{aligned}$$

As asserted, the operator $O_{\beta\to\alpha}$ preserves inner products. In particular, it preserves norms, so we have

$$\begin{aligned}
\left\| \Psi_\alpha^{(t)}(x) - O_{\beta\to\alpha} \Psi_\beta^{(t)}(y) \right\|_{\ell^2}^2 &= \left\| \Psi_\alpha^{(t)}(x) \right\|_{\ell^2}^2 + \left\| O_{\beta\to\alpha} \Psi_\beta^{(t)}(y) \right\|_{\ell^2}^2 - 2 \langle \Psi_\alpha^{(t)}(x), O_{\beta\to\alpha} \Psi_\beta^{(t)}(y) \rangle_{\ell^2} \\
&= \left\| \Psi_\alpha^{(t)}(x) \right\|_{\ell^2}^2 + \left\| \Psi_\beta^{(t)}(y) \right\|_{\ell^2}^2 - 2 \sum_{i,j\geq 1} \Psi_\alpha^{(t)}(x)[i] \, \Psi_\beta^{(t)}(y)[j] \, \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)} \\
&= D^{(t)}(x_\alpha, x_\beta).
\end{aligned}$$

Thus the operator $O_{\beta\to\alpha}$ maps the diffusion embedding $\Psi_\beta^{(t)}(X)$ into the same ℓ^2 space as the diffusion embedding $\Psi_\alpha^{(t)}(X)$, and furthermore preserves the diffusion distance between the two spaces; it is easy to see that it also preserves the diffusion distance within Γ_β. In particular, it

is possible to view both embeddings in the same ℓ^2 space, where the ℓ^2 distance is equal to the diffusion distance both within each graph Γ_α and Γ_β and between the two graphs.

Suppose now that we have three or more parameters in $\mathcal{I}$ that are of interest. Can we map all diffusion embeddings of these parameters into the same ℓ^2 space, while preserving the diffusion distances? The answer turns out to be "yes," and in fact we can use the same mapping as before. Let $\gamma \in \mathcal{I}$ be the base parameter to which all other parameters are mapped, and let $\alpha, \beta \in \mathcal{I}$ be two other arbitrary parameters. We know that we can map the embedding $\Psi_\alpha^{(t)}(X)$ into the ℓ^2 space of $\Psi_\gamma^{(t)}(X)$, and that we can also map the embedding $\Psi_\beta^{(t)}(X)$ into the ℓ^2 space of $\Psi_\gamma^{(t)}(X)$, and that these mappings will preserve diffusion distances both within Γ_γ, Γ_α, and Γ_β, and also between Γ_γ and Γ_α as well as between Γ_γ and Γ_β. We just need to show that they preserve the diffusion distance between points of Γ_α and points of Γ_β. Using essentially the same calculation as the one used to derive (17), one can obtain the following for any $v, w \in \ell^2$:

$$\langle O_{\alpha\to\gamma} v, O_{\beta\to\gamma} w\rangle_{\ell^2} = \sum_{i,j\geq 1} v[i]\, w[j]\, \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)}\rangle_{L^2(X,\mu)}.$$

But then we have:

$$\begin{aligned}
\left\| O_{\alpha\to\gamma}\Psi_\alpha^{(t)}(x) - O_{\beta\to\gamma}\Psi_\beta^{(t)}(y)\right\|_{\ell^2}^2 &= \left\| O_{\alpha\to\gamma}\Psi_\alpha^{(t)}(x)\right\|_{\ell^2}^2 + \left\| O_{\beta\to\gamma}\Psi_\beta^{(t)}(y)\right\|_{\ell^2}^2 - 2\langle O_{\alpha\to\gamma}\Psi_\alpha^{(t)}(x), O_{\beta\to\gamma}\Psi_\beta^{(t)}(y)\rangle_{\ell^2}, \\
&= \left\|\Psi_\alpha^{(t)}(x)\right\|_{\ell^2}^2 + \left\|\Psi_\beta^{(t)}(y)\right\|_{\ell^2}^2 - 2\sum_{i,j\geq 1} \Psi_\alpha^{(t)}(x)[i]\, \Psi_\beta^{(t)}(y)[j]\, \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)}\rangle_{L^2(X,\mu)} \\
&= D^{(t)}(x_\alpha, y_\beta).
\end{aligned}$$

Thus, after mapping the α and β embeddings appropriately into the γ embedding, the ℓ^2 distance is equal to all possible diffusion distances. It is therefore possible to map each of the embeddings $\{\Psi_\alpha^{(t)}(X)\}_{\alpha\in\mathcal{I}}$ into the same ℓ^2 space. In particular, one can track the evolution of the intrinsic geometry of X as it changes over $\mathcal{I}$. We summarize this discussion in the following theorem.

Theorem 3.4. *Let (X,μ) be a measure space and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ a family of kernels defined on X. Fix a parameter $\gamma \in \mathcal{I}$. If (X,μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the properties of Assumption 1, then for all $(\alpha,\beta) \in \mathcal{I}\times\mathcal{I}$,*

$$D^{(t)}(x_\alpha, y_\beta) = \left\| O_{\alpha\to\gamma}\Psi_\alpha^{(t)}(x) - O_{\beta\to\gamma}\Psi_\beta^{(t)}(y)\right\|_{\ell^2}, \quad \textit{with convergence in } L^2(X\times X, \mu\otimes\mu).$$

3.4. Historical graph

The diffusion distance $D^{(t)}(x_\alpha, y_\beta)$ defines a measure of similarity between x_α and y_β by comparing the local neighborhoods of each point in their respective graphs. The comparison is, by definition, indirect. It is possible though to use the diffusion distance to create a historical graph in which every point throughout $X \times I$ is compared directly.

Suppose, for example, that $\mathcal{I} \subset \mathbb{R}$ and that ρ is a measure for $\mathcal{I}$. Assume that $\rho(\mathcal{I}) < \infty$, $\mu(X) < \infty$, $0 < C_1 \leq k_\alpha(x,y) \leq C_2 < \infty$ for all $x, y \in X$, $\alpha \in \mathcal{I}$, and that the function $(x,y,\alpha) \mapsto k_\alpha(x,y)$ is a measurable function from $(X\times X\times\mathcal{I}, \mu\otimes\mu\otimes\rho)$ to $\mathbb{R}$. Then for each $t \in \mathbb{N}$, one can define a kernel $\overline{k}_t : (X\times\mathcal{I})\times(X\times\mathcal{I}) \to \mathbb{R}$ as

$$\overline{k}_t(x_\alpha, y_\beta) \triangleq e^{-D^{(t)}(x_\alpha, y_\beta)/\varepsilon}, \quad \text{for all } (x_\alpha, y_\beta) \in (X\times\mathcal{I})\times(X\times\mathcal{I}),$$

where $\varepsilon > 0$ is a fixed scaling parameter. The kernel $\overline{k}_t$ is a direct measure of similarity across X and the parameter space $\mathcal{I}$. Thus, when $\mathcal{I}$ is time, we think of $(X\times\mathcal{I}, \overline{k}_t)$ as defining a historical

graph in which all points throughout history are related to one another. By our assumptions, it is not hard to see that $0 < C_1(t) \leq k_t(x_\alpha, y_\beta) \leq C_2(t) < \infty$ for all $x_\alpha, y_\beta \in X \times \mathcal{I}$. Therefore we can define the density $\overline{m}_t : X \times \mathcal{I} \to \mathbb{R}$,

$$\overline{m}_t(x_\alpha) \triangleq \int_{\mathcal{I}} \int_X \overline{k}_t(x_\alpha, y_\beta)\, d\mu(y)\, d\rho(\beta), \quad \text{for all } x_\alpha \in X \times \mathcal{I},$$

as well as the normalized kernel $\overline{a}_t : (X \times \mathcal{I}) \times (X \times \mathcal{I}) \to \mathbb{R}$,

$$\overline{a}_t(x_\alpha, y_\beta) \triangleq \frac{\overline{k}_t(x_\alpha, y_\beta)}{\sqrt{\overline{m}_t(x_\alpha)}\sqrt{\overline{m}_t(y_\beta)}}, \quad \text{for all } (x_\alpha, y_\beta) \in (X \times \mathcal{I}) \times (X \times \mathcal{I}).$$

Once again using the given assumptions, one can conclude that $\overline{a}_t \in L^2(X \times \mathcal{I} \times X \times \mathcal{I}, \mu \otimes \rho \otimes \mu \otimes \rho)$. Thus it defines a Hilbert-Schmidt integral operator $\overline{A}_t : L^2(X \times \mathcal{I}, \mu \otimes \rho) \to L^2(X \times \mathcal{I}, \mu \otimes \rho)$,

$$(\overline{A}_t f)(x_\alpha) \triangleq \int_{\mathcal{I}} \int_X \overline{a}_t(x_\alpha, y_\beta) f(y_\beta)\, d\mu(y)\, d\rho(\beta), \quad \text{for all } f \in L^2(X \times \mathcal{I}, \mu \otimes \rho).$$

Let $\{\overline{\psi}_t^{(i)}\}_{i \geq 1}$ and $\{\overline{\lambda}_t^{(i)}\}_{i \geq 1}$ denote the eigenfunctions and eigenvalues of $\overline{A}_t$, respectively. The corresponding diffusion map $\overline{\Psi}_t^{(s)} : (X \times \mathcal{I}) \to \ell^2$ is given by:

$$\overline{\Psi}_t^{(s)}(x_\alpha) \triangleq \left(\left(\overline{\lambda}_t^{(i)}\right)^s \overline{\psi}_t^{(i)}(x_\alpha) \right)_{i \geq 1}, \quad \text{for all } x_\alpha \in X \times \mathcal{I}.$$

In the case when $\mathcal{I}$ is time, this diffusion map embeds the entire history of X across all of $\mathcal{I}$ into a single low dimensional space. Unlike the common embedding defined by Theorem 3.4, each point x_α is embedded in relation to the entire history of X, not just its relationship to other points y_α from the same time. As such, for each $x \in X$, one can view the trajectory of x through time as it relates to all of history, i.e., one can view:

$$T_x : \mathcal{I} \to \ell^2$$
$$T_x(\alpha) \triangleq \overline{\Psi}_t^{(s)}(x_\alpha).$$

In turn, the trajectories $\{T_x\}_{x \in X}$ can be used to define a measure of similarity between the data points in X that takes into account the history of each point.

Remark 3.5. It is also possible to define $\overline{k}_t$ in terms of the inner products of the symmetric diffusion kernels, i.e.,

$$\overline{k}_t(x_\alpha, y_\beta) \triangleq \int_X a_\alpha^{(t)}(x, u)\, a_\beta^{(t)}(y, u)\, d\mu(u).$$

Remark 3.6. The diffusion distance and corresponding analysis contained in Section 3 can be extended to the more general case in which one has a sequence of data sets $\{X_\alpha\}_{\alpha \in \mathcal{I}}$ for which there does not exist a bijective correspondence between each pair. If there is a sufficiently large set S such that $S \subset X_\alpha$ for each $\alpha \in \mathcal{I}$, then one can compute a diffusion distance from any $x_\alpha \in X_\alpha$ to any $y_\beta \in Y$ through the common set S. See Appendix A for more details.

4. Global diffusion distance

Now that we have developed a diffusion distance between pairs of data points from $(X \times \mathcal{I}) \times (X \times \mathcal{I})$, it is possible to define a global diffusion distance between Γ_α and Γ_β. The aim here is to define a diffusion distance that gives a global measure of the change in X from α to β. In turn, when applied over the whole parameter space, one can organize the global behavior of the data as it changes over $\mathcal{I}$. For each diffusion time $t \in \mathbb{N}$, let $\mathcal{D}^{(t)} : \{\Gamma_\alpha\}_{\alpha\in\mathcal{I}} \times \{\Gamma_\alpha\}_{\alpha\in\mathcal{I}} \to \mathbb{R}$ be this global diffusion distance, where

$$\begin{aligned}\mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta)^2 &\triangleq \left\|A_\alpha^t - A_\beta^t\right\|_{HS}^2 \\ &= \left\|a_\alpha^{(t)} - a_\beta^{(t)}\right\|_{L^2(X\times X, \mu\otimes\mu)}^2 \\ &= \iint_{X\times X} \left(a_\alpha^{(t)}(x,y) - a_\beta^{(t)}(x,y)\right)^2 \, d\mu(x)\, d\mu(y).\end{aligned}$$

In fact, since μ is a σ-finite measure, the global diffusion distance can be written in terms of the pointwise diffusion distance by applying Tonelli's Theorem:

$$\mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta)^2 = \int_X D^{(t)}(x_\alpha, x_\beta)^2 \, d\mu(x).$$

Thus the global diffusion distance measures the similarity between Γ_α and Γ_β by comparing the behavior of each of the corresponding diffusions on each of the graphs. Therefore, the global diffusion distance will be small if Γ_α and Γ_β have similar geometry, and large if their geometry is significantly different.

As with the pointwise diffusion distance $D^{(t)}$, the global diffusion distance can be written in a simplified form in terms of the spectral decompositions of the operators A_α and A_β.

Theorem 4.1. *Let (X, μ) be a measure space and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ a family of kernels defined on X. If (X, μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the properties of Assumption 1, then the global diffusion distance at time t between Γ_α and Γ_β can be written as:*

$$\mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta)^2 = \sum_{i,j\geq 1} \left(\left(\lambda_\alpha^{(i)}\right)^t - \left(\lambda_\beta^{(j)}\right)^t\right)^2 \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)}^2. \tag{18}$$

We postpone the proof till the end of this section. Equation (18) gives a new way to interpret the global diffusion graph distance. The orthonormal basis $\{\psi_\alpha^{(i)}\}_{i\geq 1}$ is a set of diffusion coordinates for Γ_α, while the orthonormal basis $\{\psi_\beta^{(j)}\}_{j\geq 1}$ is a set of diffusion coordinates for Γ_β. Interpreting the summands of (18) in this context, we see that the global diffusion distance measures the similarity of Γ_α and Γ_β by taking a weighted rotation of one coordinate system into the other.

Remark 4.2. As with the pointwise diffusion distance, the asymptotic behavior of the global diffusion distance when $\{\Gamma_\alpha\}_{\alpha\in\mathcal{I}}$ is a family of connected graphs is both interesting and easy to characterize. Under the same connectivity assumptions as Remark 3.3, it is not hard to show that

$$\lim_{t\to\infty} \mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta)^2 = 2\left(1 - \left\langle \sqrt{m_\alpha(\cdot)}, \sqrt{m_\beta(\cdot)} \right\rangle_{L^2(X,\mu)}^2\right).$$

Proof of Theorem 4.1. Since

$$\mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta)^2 = \int_X D_t(x_\alpha, x_\beta)^2 \, d\mu(x),$$

we can build upon Theorem 3.2. In particular, we have

$$\mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta)^2 = \int_X \left(\sum_{i\geq 1} \left(\lambda_\alpha^{(i)}\right)^{2t} \psi_\alpha^{(i)}(x)^2 + \sum_{j\geq 1} \left(\lambda_\beta^{(j)}\right)^{2t} \psi_\beta^{(j)}(x)^2 \right.$$
$$\left. -2 \sum_{i,j\geq 1} \left(\lambda_\alpha^{(i)}\right)^t \left(\lambda_\beta^{(j)}\right)^t \psi_\alpha^{(i)}(x)\, \psi_\beta^{(j)}(x)\, \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)} \right) d\mu(x).$$

As in the proof of Theorem 3.2 we have three terms that we shall evaluate separately. Focusing on the cross terms as before, we would like to switch the integral and the summation; this time we need to show

$$\sum_{i,j\geq 1} \int_X \left| \left(\lambda_\alpha^{(i)}\right)^t \left(\lambda_\beta^{(j)}\right)^t \psi_\alpha^{(i)}(x)\, \psi_\beta^{(j)}(x)\, \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)} \right| d\mu(x) < \infty. \tag{19}$$

One can show (19) by using Hölder's Theorem, the Cauchy-Schwarz inquality, and the assumption that A_α is a trace class operator for each $\alpha \in \mathcal{I}$. Therefore we can switch the integral and the summation, which gives:

$$\int_X \sum_{i,j\geq 1} \left(\lambda_\alpha^{(i)}\right)^t \left(\lambda_\beta^{(j)}\right)^t \psi_\alpha^{(i)}(x)\, \psi_\beta^{(j)}(x)\, \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)} \, d\mu(x) = \sum_{i,j\geq 1} \left(\lambda_\alpha^{(i)}\right)^t \left(\lambda_\beta^{(j)}\right)^t \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)}^2. \tag{20}$$

A similar calculation also shows that for each $\alpha \in \mathcal{I}$,

$$\int_X \sum_{i\geq 1} \left(\lambda_\alpha^{(i)}\right)^{2t} \psi_\alpha^{(i)}(x)^2 \, d\mu(x) = \sum_{i\geq 1} \left(\lambda_\alpha^{(i)}\right)^{2t}. \tag{21}$$

Putting (20) and (21) together, we arrive at:

$$\mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta)^2 = \sum_{i\geq 1} \left(\lambda_\alpha^{(i)}\right)^{2t} + \sum_{j\geq 1} \left(\lambda_\beta^{(j)}\right)^{2t} - 2 \sum_{i,j\geq 1} \left(\lambda_\alpha^{(i)}\right)^t \left(\lambda_\beta^{(j)}\right)^t \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)}^2. \tag{22}$$

Furthermore, by the Spectral Theorem we can take $\{\psi_\alpha^{(i)}\}_{i\geq 1}$ and $\{\psi_\beta^{(j)}\}_{j\geq 1}$ to be orthonormal bases for $L^2(X,\mu)$. In particular,

$$\sum_{i\geq 1} \langle \psi_\alpha^{(i)}, \psi_\beta^{(j_0)} \rangle^2 = \sum_{j\geq 1} \langle \psi_\alpha^{(i_0)}, \psi_\beta^{(j)} \rangle^2 = 1, \quad \text{for all } i_0, j_0 \geq 1.$$

Therefore we can simplify (22) to

$$\mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta)^2 = \sum_{i,j\geq 1} \left(\left(\lambda_\alpha^{(i)}\right)^t - \left(\lambda_\beta^{(j)}\right)^t \right)^2 \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)}^2.$$

□

5. Random sampling theorems

In applications, the given data is finite and often times sampled from some continuous data set X. In this section we examine the behavior of the pointwise and global diffusion distances when applied to a randomly sampled, finite collection of samples taken from X.

5.1. Updated assumptions

In order to frame this discussion in the appropriate setting, we update our assumptions on the measure space (X, μ) and the kernels $\{k_\alpha\}_{\alpha\in\mathcal{I}}$. The results from this section will rely heavily upon the work contained in [14, 15], and so we follow their lead. First, for any $l \in \mathbb{N}$, let $C_b^l(X)$ denote the set of continuous bounded functions on X such that all derivatives of order l exist and are themselves continuous, bounded functions.

Assumption 2. *We assume the following properties:*

1. *The measure μ is a probability measure, so that $\mu(X) = 1$.*
2. *X is a bounded open subset of $\mathbb{R}^d$ that satisfies the cone condition (see page 93 of [16]).*
3. *For each $\alpha \in \mathcal{I}$, the kernel k_α is symmetric, positive definite, and bounded from above and below, so that*
$$0 < C_1(\alpha) \leq k_\alpha(x, y) \leq C_2(\alpha) < \infty.$$
4. *For each $\alpha \in \mathcal{I}$, $k_\alpha \in C_b^{d+1}(X \times X)$.*

Note that every property from Assumption 1 is either contained in or can be derived from the properties in Assumption 2. Therefore the results of the previous sections still apply under these new assumptions.

The first assumption that μ be a probability measure is needed since we will be randomly sampling points from X. The probability measure from which we sample is μ. The second and fourth assumptions are necessary to apply certain Sobolev embedding theorems which are integral to constructing a reproducing kernel Hilbert space that contains the family of kernels $\{a_\alpha\}_{\alpha\in\mathcal{I}}$ and their empirical equivalents. More details can be found in Appendix B.

5.2. Sampling and finite graphs

Consider the space X and suppose that $X_n \triangleq \{x^{(1)}, \ldots, x^{(n)}\} \subset X$ are sampled i.i.d. according to μ. We are going to discretize the framework we have developed to accommodate the samples X_n. Let $\Gamma_{\alpha,n} \triangleq (X_n, k_\alpha|_{X_n})$ be the finite graph with vertices X_n and weighted edges given by $k_\alpha|_{X_n}$. We now define the finite, matrix equivalents to the continuous operators from section 3.1. To start, first define for each $\alpha \in \mathcal{I}$ the $n \times n$ matrices $\mathbb{K}_\alpha$ as:

$$\mathbb{K}_\alpha[i, j] \triangleq \frac{1}{n} k_\alpha(x^{(i)}, x^{(j)}), \quad \text{for all } i, j = 1, \ldots, n.$$

We also define the corresponding diagonal degree matrices $\mathbb{D}_\alpha$ as:

$$\mathbb{D}_\alpha[i, i] \triangleq \frac{1}{n} \sum_{j=1}^{n} k_\alpha(x^{(i)}, x^{(j)}) = \sum_{j=1}^{n} \mathbb{K}_\alpha[i, j], \quad \text{for all } i = 1, \ldots, n.$$

Finally, the discrete analog of the operator A_α is given by the matrix $\mathbb{A}_\alpha$, which is defined as

$$\mathbb{A}_\alpha \triangleq \mathbb{D}_\alpha^{-\frac{1}{2}} \mathbb{K}_\alpha \mathbb{D}_\alpha^{-\frac{1}{2}}, \quad \text{for all } \alpha \in \mathcal{I}.$$

We can now define the pointwise and global diffusion distances for the finite graphs $\{\Gamma_{\alpha,n}\}_{\alpha\in\mathcal{I}}$ in terms of the matrices $\{\mathbb{A}_\alpha\}_{\alpha\in\mathcal{I}}$. Set $x_\alpha^{(i)} \triangleq (x^{(i)},\alpha) \in X_n\times\mathcal{I}$, and let $D_n^{(t)} : (X_n\times\mathcal{I})\times(X_n\times\mathcal{I}) \to \mathbb{R}$ denote the empirical version of the pointwise diffusion distance. We define it as:

$$\begin{aligned} D_n^{(t)}(x_\alpha^{(i)}, x_\beta^{(j)})^2 &\triangleq n^2 \left\|\mathbb{A}_\alpha^t[i,\cdot] - \mathbb{A}_\beta^t[j,\cdot]\right\|_{\mathbb{R}^n}^2 \\ &= n^2 \sum_{k=1}^{n} \left(\mathbb{A}_\alpha^t[i,k] - \mathbb{A}_\beta^t[j,k]\right)^2. \end{aligned}$$

Let $\mathcal{D}_n^{(t)} : \{\Gamma_{\alpha,n}\}_{\alpha\in\mathcal{I}} \times \{\Gamma_{\alpha,n}\}_{\alpha\in I} \to \mathbb{R}$ denote the empirical global diffusion distance, where

$$\begin{aligned} \mathcal{D}_n^{(t)}(\Gamma_{\alpha,n}, \Gamma_{\beta,n})^2 &\triangleq \left\|\mathbb{A}_\alpha^t - \mathbb{A}_\beta^t\right\|_{HS} \\ &= \sum_{i,j=1}^{n} \left(\mathbb{A}_\alpha^t[i,j] - \mathbb{A}_\beta^t[i,j]\right)^2. \end{aligned}$$

We then have the following two theorems relating $D_n^{(t)}$ to D_t and $\mathcal{D}_n^{(t)}$ to $\mathcal{D}^{(t)}$, respectively.

Theorem 5.1. *Suppose that (X,μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the conditions of Assumption 2. Let $n \in \mathbb{N}$ and sample $X_n = \{x^{(1)},\dots,x^{(n)}\} \subset X$ i.i.d. according to μ; also let $t \in \mathbb{N}$, $\tau > 0$, and $\alpha,\beta \in \mathcal{I}$. Then, with probability $1 - 2e^{-\tau}$,*

$$\left|D^{(t)}(x_\alpha^{(i)}, x_\beta^{(j)}) - D_n^{(t)}(x_\alpha^{(i)}, x_\beta^{(j)})\right| \le C(\alpha,\beta,d,t)\frac{\sqrt{\tau}}{\sqrt{n}}, \quad \text{for all } i,j = 1,\dots,n.$$

Theorem 5.2. *Suppose that (X,μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the conditions of Assumption 2. Let $n \in \mathbb{N}$ and sample $X_n = \{x^{(1)},\dots,x^{(n)}\} \subset X$ i.i.d. according to μ; also let $t \in \mathbb{N}$, $\tau > 0$, and $\alpha,\beta \in \mathcal{I}$. Then, with probability $1 - 2e^{-\tau}$,*

$$\left|\mathcal{D}^{(t)}(\Gamma_\alpha,\Gamma_\beta) - \mathcal{D}_n^{(t)}(\Gamma_{\alpha,n},\Gamma_{\beta,n})\right| \le C(\alpha,\beta,d,t)\frac{\sqrt{\tau}}{\sqrt{n}}.$$

6. Applications

6.1. Change detection in hyperspectral imagery data

In this section we consider the problem of change detection in hyperspectral imagery (HSI) data. The main ideas are the following. A hyperspectral image can be thought of as a data cube C, with dimensions $L\times W\times D$. The cube C corresponds to an image whose pixel dimensions are $L\times W$. A hyperspectral camera measures the reflectance of this image at D different wavelengths, giving one D images, which, put together, give one the cube C. Thus we think of a hyperspectral image as a regular image, but, each pixel now has a spectral signature in $\mathbb{R}^D$.

The change detection problem is the following. Suppose you have one scene for which you have several hyperspectral images taken at different times. These images can be taken under different weather conditions, lighting conditions, during different seasons of the year, and even with different cameras. The goal is to determine what has changed from one image to the next.

To test the diffusion distance in this setting, we used some of the data collected in [1]. Using a hyperspectral camera that captured 124 different wavelengths, the authors of [1] collected hyperspectral images of a particular scene during August, September, October, and November (one

(a) August (b) September (c) October (d) November

Figure 1: Color images of the four months.

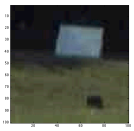


Figure 2: Color image of the changed scene (taken in October).

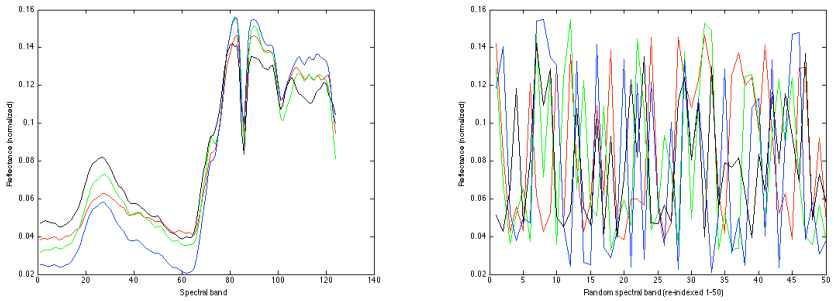
image for each month). In October, they also recorded a fifth image in which they introduced two small changes into the scene as a means for testing change detection algorithms. For our purposes, we selected a particular $100 \times 100 \times 124$ sub-cube across all five images that contains one of the aforementioned introduced changes. Color images of the four months are given in figure 1, while the fifth image taken in October and containing the change is given in figure 2. In all five images one can see in the foreground grass and in the background a tree line, with a metal panel resting on the grass. In the additional fifth image, there is also a small tarp sitting on the grass. The images were obviously taken during different times of the year, ranging from Summer to Fall, and it is also evident that the lighting is different from image to image. One can see these changes in how the spectral signature of a particular pixel changes from month to month; see figure 3(a) for an example.

The authors did use the same camera for each image though, so in order to simulate one using different cameras we did the following. For each of the five images, we randomly selected 50 of the 124 bands to use; we also randomly reordered each set of 50 bands. To see an example of these new spectra, we refer the reader to figure 3(b). While the seasonal and lighting changes affected the spectra, it is clear from figure 3(a) that they are still at least intuitively comparable since the same camera was used. Using the five random "cameras," however, makes a direct comparison between months now much more difficult.

We set the parameter space as $\mathcal{I} = \{\text{aug, sep, oct, nov, chg}\}$, where chg denotes the data set with the change in it. We also set $\mathcal{I}^{(4)} \triangleq \{\text{aug, sep, oct, nov}\} \subset \mathcal{I}$. For each $\alpha \in \mathcal{I}$, we let X_α denote the corresponding $100 \times 100 \times 50$ hyperspectral image taken with our "random" camera (derived from the original hyperspectral images as described above). For each month as well as the changed data set, we computed a Gaussian kernel of the form:

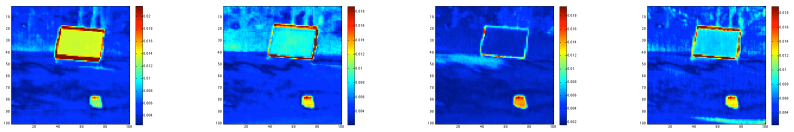
$$k_\alpha(x, y) = e^{-\|x-y\|^2/\varepsilon(\alpha)^2}, \quad \text{for all } \alpha \in \mathcal{I}, \;\; x, y \in X_\alpha,$$

where $\varepsilon(\alpha)$ was selected so that the corresponding symmetric diffusion operator (matrix) A_α would have second eigenvalue $\lambda_\alpha^{(2)} \approx 0.97$. By forcing each diffusion operator to have approximately the same second eigenvalue, the five diffusion processes will spread at approximately the same rate. We then computed the diffusion distance between a pixel x taken from X_{chg} and its



(a) Original spectra (b) Random "camera" spectra

Figure 3: Spectrum of a single pixel across the four months. Red: August, green: September, blue: October, black: November.

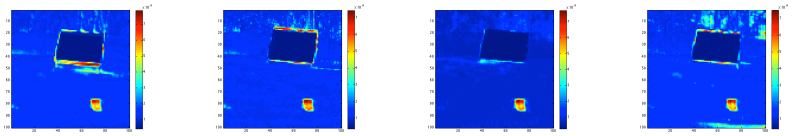


(a) August (b) September (c) October (d) November

Figure 4: Map of $D^{(1)}(x_{\text{chg}}, x_\alpha)$ for each $\alpha \in \mathcal{I}^{(4)}$

corresponding pixel in X_α for each $\alpha \in \mathcal{I}^{(4)}$, i.e., we computed $D^{(t)}(x_{\text{chg}}, x_\alpha)$. The results for $t = 1$ are given in figure 4, while the asymptotic diffusion distance as $t \to \infty$ is given in figure 5.

For the diffusion time $t = 1$, we see from the maps in figure 4 that the tarp is recognized as a change. However, other changes due to the lighting or the change in seasons also appear. For example, even in October, the small change in the shadow is visible, while in August, September, and November the change in lighting causes the panel to be highlighted. Also, in some months even the trees have a weak, but noticeable difference in the their diffusion distances. When we allow $t \to \infty$ though, the smaller clusters merge together and the changes due to lighting and seasonal differences are filtered out. As one can see from figure 5, all that is left is the change due to the added tarp (note that the change around the border of the panel is due to it being slightly shifted from month to month). Thus we see in practice that the diffusion distance can be



(a) August (b) September (c) October (d) November

Figure 5: Map of $\lim_{t\to\infty} D^{(t)}(x_{\text{chg}}, x_\alpha)$ for each $\alpha \in \mathcal{I}^{(4)}$

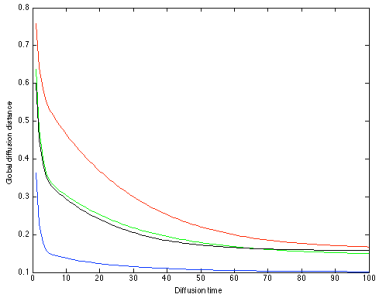


Figure 6: Global diffusion distance. Red: $\mathcal{D}^{(t)}(\Gamma_{\text{chg}}, \Gamma_{\text{aug}})$, green: $\mathcal{D}^{(t)}(\Gamma_{\text{chg}}, \Gamma_{\text{sep}})$, blue: $\mathcal{D}^{(t)}(\Gamma_{\text{chg}}, \Gamma_{\text{oct}})$, black: $\mathcal{D}^{(t)}(\Gamma_{\text{chg}}, \Gamma_{\text{nov}})$

used to filter types of changes at different scales.

We also computed the global diffusion distances between the changed data set and the four months. The results are given in figure 6. We see several intuitions borne out in this particular application. First, the closer the month in real time to October (the month in which the changed data set was recorded), the smaller the global diffusion distance. Secondly, we see that as the diffusion time t gets larger, the smaller the global diffusion distance.

6.2. Parameterized difference equations

Suppose one has the following example developed by Mezić and Lederman of a dynamical system that depends on a certain input parameter. We take the data set X_α to be the set of orbits for the parameter α. If one can define a suitable set of kernels $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ on the orbits, then using the diffusion distance it is possible to not only organize the behavior of the system for a fixed parameter, but to also see how the dynamics of the system change as as the parameter is changed.

Consider the standard map, which is an area preserving chaotic map from the torus $\mathbb{T}^2 \triangleq 2\pi(S^1 \times S^1)$ onto itself. It is defined as:

$$p_{\ell+1} \triangleq p_\ell + \alpha \sin(\theta_\ell),$$
$$\theta_{\ell+1} \triangleq \theta_\ell + p_{\ell+1},$$

where $\alpha \in \mathcal{I} = [0, \infty)$ is a parameter, $\ell \in \mathbb{N}\cup\{0\}$, and $(p_\ell, \theta_\ell) \in \mathbb{T}^2$. When $\alpha = 0$, the map consists solely of periodic and quasiperiodic orbits. For $\alpha > 0$, the map is is increasingly nonlinear as α grows, which in turn increases the chances of observing chaotic dynamics.

Using the ideas developed in [17, 18], it is possible to define a kernel k_α on the orbits of the standard map with parameter α. One can in turn use this kernel to define a diffusion map on these orbits; an example for a small α is given in figure 7. Using the ideas contained in Section 3, it is also possible to embed each diffusion map, for all $\alpha \in \mathcal{I}$, into a single embedding. Doing so allows one to observe how the dynamics of the system change as the parameter is increased; see figure 8 for more details. In the forthcoming paper [19], we give a full treatment of these ideas.

6.3. Global embeddings

Another application of the global diffusion distance is that it gives a metric by which to compute a "graph of graphs". By this we mean the following: given our family of graphs $\{\Gamma_\alpha\}_{\alpha\in\mathcal{I}}$,

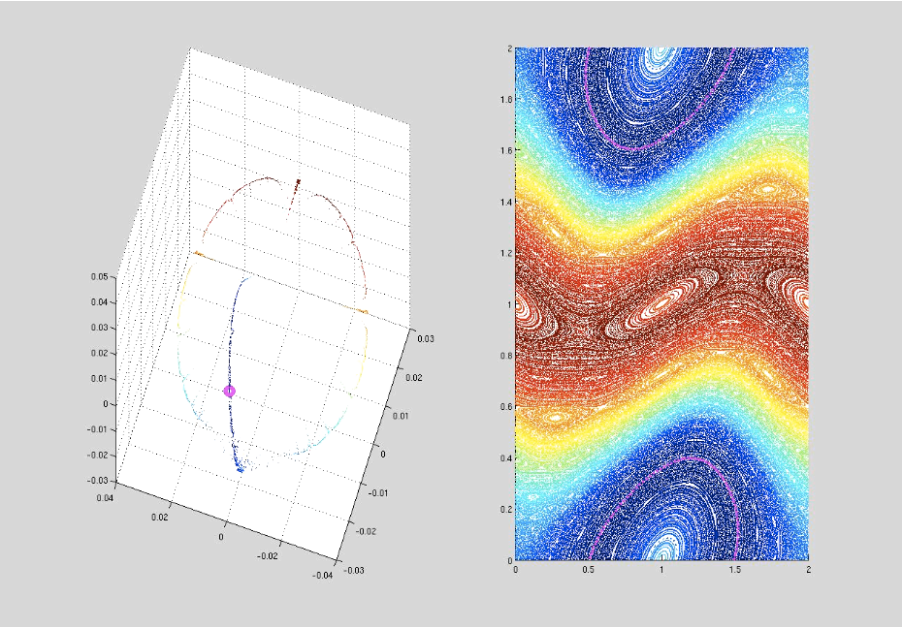


Figure 7: Diffusion map of the orbits of the standard map for a small α. The color of the embedded point on the left corresponds to the orbit of the same color on the right. A particular embedded point and orbit are highlighted in purple

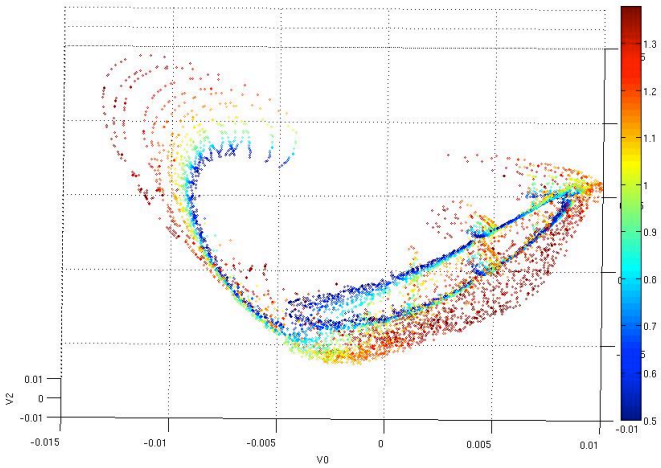


Figure 8: Common diffusion embedding of the orbits of the standard map across several values of α. The color of the embedded point indicates the value of α used in the standard map. Notice, in particular, that many of the periodic and quasiperiodic orbits for low values of α that are embedded into the central ring of the embedding turn into chaotic orbits for higher values of α. This in turn is realized by the diffusion map as the embedding has less structure.

we can compute a new graph $\Omega_t \triangleq (\{\Gamma_\alpha\}_{\alpha\in\mathcal{I}}, \overline{k}_t)$, in which the vertices of Ω_t are the graphs $\{\Gamma_\alpha\}_{\alpha\in\mathcal{I}}$ and the kernel $\overline{k}_t : \{\Gamma_\alpha\}_{\alpha\in\mathcal{I}} \times \{\Gamma_\alpha\}_{\alpha\in\mathcal{I}} \to \mathbb{R}$ is a function of the global diffusion distance $\mathcal{D}^{(t)}$. One natural way to define $\overline{k}_t$ is via Gaussian weights:

$$\overline{k}_t(\Gamma_\alpha, \Gamma_\beta) \triangleq e^{-\mathcal{D}^{(t)}(\Gamma_\alpha,\Gamma_\beta)^2/\varepsilon^2}, \quad \text{for all } \alpha,\beta \in \mathcal{I}. \tag{23}$$

Note that for each diffusion time t, we have different kernel $\overline{k}_t$ which results in a different graph Ω_t. Fixing a specific, but arbitrary diffusion time t, one can in turn construct a new diffusion operator on the graph Ω_t by using $\overline{k}_t$ as the underlying kernel. For example, if $\mathcal{I}$ is finite and we let $\overline{m}_t : \{\Gamma_\alpha\}_{\alpha\in\mathcal{I}} \to \mathbb{R}$ be the density of $\overline{k}_t$, where

$$\overline{m}_t(\Gamma_\alpha) \triangleq \sum_{\beta\in\mathcal{I}} \overline{k}_t(\Gamma_\alpha, \Gamma_\beta),$$

then the corresponding symmetric diffusion kernel $\overline{a}_t : \{\Gamma_\alpha\}_{\alpha\in\mathcal{I}} \times \{\Gamma_\alpha\}_{\alpha\in\mathcal{I}} \to \mathbb{R}$ would be defined as

$$\overline{a}_t(\Gamma_\alpha, \Gamma_\beta) \triangleq \frac{\overline{k}_t(\Gamma_\alpha, \Gamma_\beta)}{\sqrt{\overline{m}_t(\Gamma_\alpha)}\,\sqrt{\overline{m}_t(\Gamma_\beta)}}.$$

Since we are assuming $\mathcal{I}$ is finite, one can think of $\overline{a}_t$ as a $|\mathcal{I}| \times |\mathcal{I}|$ matrix, and one can compute the eigenvectors and eigenvalues of $\overline{a}_t$. The standard diffusion embedding can then be used to cluster the family of graphs $\{\Gamma_\alpha\}_{\alpha\in\mathcal{I}}$. Keep in mind that this second diffusion embedding has its own diffusion time, completely separate from the diffusion time t on the individual graphs $\{\Gamma_\alpha\}_{\alpha\in\mathcal{I}}$. Finally, if $\mathcal{I}$ is infinite, then one must define an appropriate measure on the family of graphs $\{\Gamma_\alpha\}_{\alpha\in\mathcal{I}}$ or on the parameter space $\mathcal{I}$.

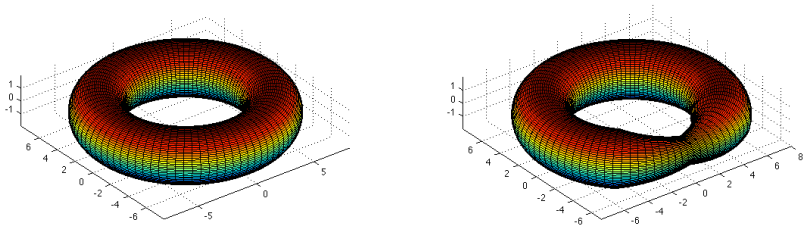
We apply this idea to the following example. Our initial data set X is a torus with a central radius of six and a lateral radius of two; more formally, $X \triangleq (6S^1) \times (2S^1)$, where S^1 is the unit circle in $\mathbb{R}^2$. An image of X is given in figure 9(a). We assume that the central circle $6S^1$ and the lateral circle $2S^1$ are oriented, so that each point on the torus has a specific coordinate location (note that while $X \subset \mathbb{R}^3$, the points of the torus have a two dimensional coordinate system consisting of two angles, one for the central circle and one for the lateral circle).

From X we build a family of "pinched" torii as follows. We pick an angle on the central circle $6S^1$, say θ_0, and we pinch the torus at θ_0 so that its lateral radius at this angle is now r_0, where $r_0 < 2$. So that we do not rip the torus, from a starting angle θ_s, the lateral radius will decrease linearly from 2 at θ_s to r_0 at θ_0, and then increase linearly from r_0 at θ_0 back to 2 at some ending angle θ_e. The lateral radius of this new torus will be 2 at all other angles on the central circle. For an example see figure 9(b).

We create several pinched torii as follows. We take three different angles to pinch the torus at: $\theta_0 = \pi/2, \pi$, and $3\pi/2$. At each of these three angles, we pinch the torus so that the lateral radius r_0 at θ_0 can take one of ten values: $r_0 = 1, 1.1, 1.2, \ldots, 1.9$. The starting and ending angles for each pinch are offset from θ_0 by $\pi/4$ radians, so that $\theta_s = \theta_0 - \pi/4$ and $\theta_e = \theta_0 + \pi/4$. Thus we have 30 different pinched torii, which along with the original torus, gives us a family of 31 torii.

Our goal is to build a graph in which each vertex is one the 31 torii. To do so we approximate the global diffusion distance between each pair of torii by taking 7744 random samples from X (using the uniform distribution), and then using the same corresponding samples for each pinched torus. For each torus we used a Gaussian kernel of the form

$$k_\alpha(x, y) = e^{-\|x-y\|^2/\varepsilon(\alpha)^2}, \quad \text{for all } \alpha = (\theta_0, r_0),$$



(a) Regular torus (b) Pinched torus

Figure 9: Regular and pinched torii

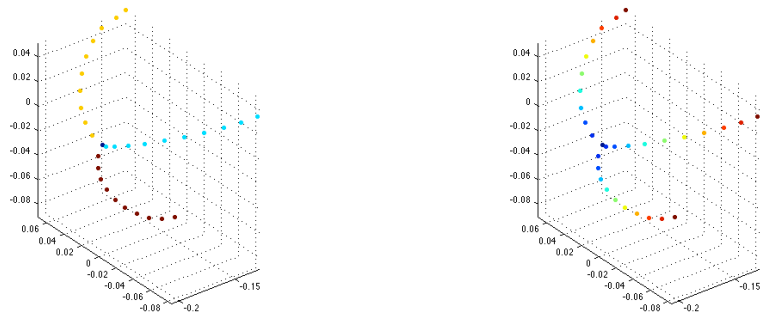
where $\varepsilon(\alpha)$ was selected so that the corresponding symmetric diffusion operator (matrix) A_α would have second eigenvalue $\lambda_\alpha^{(2)} = 0.5$. The pairwise global diffusion distance was further approximated by taking the top ten eigenvalues and eigenvectors of each of the 31 diffusion operators, and was then computed for diffusion time $t = 2$ using Theorem 4.1. Two remarks: first, the diffusion time $t = 2 = 1/(1 - \lambda_\alpha^{(2)})$ corresponds to the approximate time it would take for the diffusion process to spread through each of the graphs; secondly, by Theorem 5.2, this approximate global diffusion distance is, with high probability, nearly equal to the true global diffusion distance between each of the torii.

After computing the pairwise global diffusion distances, we constructed the kernel $\overline{k}_t$, for $t = 2$, defined in equation (23). We took ε in this kernel to be the median of all pairwise global diffusion distances between the 31 torii. We then computed the symmetric diffusion operator for this graph of graphs, which turned out to have second eigenvalue $\lambda^{(2)} \approx 0.48$. We took the top three eigenvalues and eigenvectors of the diffusion operator, and used them to compute the diffusion map into $\mathbb{R}^3$ at diffusion time $t' \approx 1/(1 - 0.48) = 1.92$.

A plot of this diffusion map is given in figure 10. The central, dark blue, circle corresponds to the regular torus in both images. In figure 10(a), the other three colors correspond to the angle at which the torus was pinched. In figure 10(b), the colors correspond to the strength of the pinch (dark blue - no pinch, dark red - strongest pinch). As one can see, the diffusion embedding organizes the torii by both the location of the pinch (i.e. what arc the embedded torus lies on), and the strength of the pinch (i.e. how far from the regular torus each pinched torus lies), giving a global view of how the data set changes over the parameter space.

7. Conclusion

In this paper we have generalized the diffusion distance to work on a changing graph. This new distance, along with the corresponding diffusion maps, allow one to understand how the intrinsic geometry of the data set changes over the parameter space. We have also defined a global diffusion distance between graphs, and used this to construct meta graphs in which each vertex of the meta graph corresponds to a graph. Formulas for each of these diffusion distances in terms of the spectral decompositions of the relevant diffusion operators have been proven, giving a simple and efficient way to approximate these diffusion distances. Finally, it was shown



(a) Colored by location of pinch

(b) Colored by strength of pinch

Figure 10: Diffusion embedding of the 31 torii

that a random, finite sample of data points from a continuous, changing data set X is, with high probability, enough to approximate the diffusion distance and the global diffusion distance to high accuracy.

Future work could include generalizing these notions of diffusion distance further so that they can apply the sequences of graphs in which vertices are added and dropped (i.e. in which there is no bijective correspondence between graphs). Also, it would be interesting to investigate how this work fits in with the recent research on vectorized diffusion operators contained in [20, 21]

8. Acknowledgements

This research was supported by Air Force Office of Scientific Research STTR FA9550-10-C-0134 and by Army Research Office MURI W911NF-09-1-0383.

Appendix A. Non-bijective correspondence

In this appendix we consider the case in which our changing data set does not have a single bijective correspondence across the parameter set $\mathcal{I}$. We make a few small changes to the notation. Continue to let $\mathcal{I}$ denote the parameter space, but let $(\mathcal{X}, \mu)$ denote a "global" measure space (with $\mathcal{X} \subseteq \mathbb{R}^d$). Our changing data is given by $\{X_\alpha\}_{\alpha\in\mathcal{I}}$ with data points $x_\alpha \in X_\alpha$, and satisfies

$$X_\alpha \subseteq \mathcal{X}, \quad \text{for all } \alpha \in \mathcal{I}.$$

We assume that each data set X_α is a measurable set under μ. Suppose, additionally, that there exists a sufficiently large set $S \subset \mathcal{X}$ such that

$$S \subset X_\alpha, \quad \text{for all } \alpha \in \mathcal{I}.$$

We maintain the remaining notations and assumptions from Section 3, and simply update them to apply for each X_α. In particular, for each $\alpha \in \mathcal{I}$, we have the symmetric diffusion kernel $a_\alpha : X_\alpha \times X_\alpha \to \mathbb{R}$, with corresponding trace class operator $A_\alpha : L^2(X_\alpha, \mu) \to L^2(X_\alpha, \mu)$. The set of functions $\{\psi_\alpha^{(i)}\}_{i\geq 1} \subset L^2(X_\alpha, \mu)$ still denote a set of orthonormal eigenfunctions for A_α, with

corresponding eigenvalues $\{\lambda_\alpha^{(i)}\}_{i\geq 1}$. The diffusion map is still given by $\Psi_\alpha^{(t)} : X_\alpha \to \ell^2$, with $\Psi_\alpha^{(t)}(x_\alpha) = \left(\left(\lambda_\alpha^{(i)}\right)^t \psi_\alpha^{(i)}(x_\alpha)\right)_{i\geq 1}$.

Under this more general setup, for any $\alpha, \beta \in \mathcal{I}$, the sets $X_\alpha \setminus X_\beta$ and $X_\beta \setminus X_\alpha$ may be nonempty. Thus it is not possible to compare the diffusions on Γ_α and Γ_β as they spread through each graph. On the other hand, since we have a common set $S \subset X_\alpha \cap X_\beta$, we can compare the diffusion centered at $x_\alpha \in X_\alpha$ with the diffusion centered at $y_\beta \in X_\beta$ as they spread through the subgraphs of Γ_α and Γ_β with common vertices S. Formally, we define this diffusion distance as:

$$D^{(t)}(x_\alpha, y_\beta; S)^2 \triangleq \int_S \left(a_\alpha^{(t)}(x_\alpha, s) - a_\beta^{(t)}(y_\beta, s)\right)^2 d\mu(s), \quad \text{for all } \alpha, \beta \in \mathcal{I}, \ (x_\alpha, y_\beta) \in X_\alpha \times X_\beta.$$

A result similar to Theorem 3.4 can be had for this subgraph diffusion distance. Since the eigenfunctions for A_α will not be orthonormal when restricted to $L^2(S, \mu)$, one must use an additional orthonormal basis $\{e^{(i)}\}_{i\geq 1}$ for $L^2(S, \mu)$ when rotating the diffusion maps across $\mathcal{I}$ into a common embedding. In particular, we define a new family of rotation maps $O_{\alpha,S} : \ell^2 \to \ell^2$ as:

$$O_{\alpha,S}\, v \triangleq \left(\sum_{j\geq 1} v[j]\, \langle e^{(i)}, \psi_\alpha^{(j)} \rangle_{L^2(S,\mu)}\right)_{i\geq 1}.$$

Using these rotation maps, along with the same ideas from Section 3, one can show:

$$D^{(t)}(x_\alpha, y_\beta; S) = \left\| O_{\alpha,S} \Psi_\alpha^{(t)}(x_\alpha) - O_{\beta,S} \Psi_\beta^{(t)}(y_\beta) \right\|_{\ell^2}, \quad \text{with convergence in } L^2(X_\alpha \times X_\beta, \mu \otimes \mu).$$

Appendix B. Proof of random sampling theorems

In this appendix we prove the random sampling Theorems 5.1 and 5.2 from section 5. Throughout the appendix we shall assume that (X, μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy Assumption 2.

The proof shall rely upon a result from [14] as well as several results on the asymmetric graph Laplacian $I - P$ that are contained in [15]. All of these results are easily translated for our family of operators $\{A_\alpha\}_{\alpha\in\mathcal{I}}$, and we shall simply restate the needed results from [15] in these terms.

Appendix B.1. Reproducing kernel Hilbert spaces

Critical to our analysis will the be existence of a single reproducing kernel Hilbert space (RKHS) that contains the set of kernels $\{a_\alpha\}_{\alpha\in\mathcal{I}}$, their empirical approximations, and related functions. In [15] such a RKHS is constructed. Here we recall the definition of a RKHS as well the aforementioned construction.

A set $\mathcal{H}$ is a RKHS if it is a Hilbert space of functions $f : X \to \mathbb{R}$ such that for each $x \in X$, there exists a constant $C(x)$ so that

$$f(x) \leq C(x)\, \|f\|_{\mathcal{H}}.$$

The name RKHS comes from the fact that one can show that there is a unique symmetric, positive definite kernel $h : X \times X \to \mathbb{R}$ associated with $\mathcal{H}$ such that for each $f \in \mathcal{H}$,

$$f(x) = \langle f, h(x, \cdot)\rangle_{\mathcal{H}}, \quad \text{for all } x \in X.$$

We utilize a specific RKHS first presented in [15]; the construction is rewritten here for completeness. Let l be a positive integer, and define the Sobolev space $\mathcal{H}^l$ as

$$\mathcal{H}^l \triangleq \{f \in L^2(X, dx) : D^\gamma f \in L^2(X, dx) \text{ for all } |\gamma| = l\},$$

where $D^\gamma f$ is the weak derivative of f with respect to the multi-index $\gamma \triangleq (\gamma_1, \ldots, \gamma_d) \in \mathbb{N}^d$, $|\gamma| \triangleq \gamma_1 + \cdots + \gamma_d$, and dx denotes the Lebesgue measure. The space $\mathcal{H}^l$ is a separable Hilbert space with scalar product

$$\langle f, g\rangle_{\mathcal{H}^l} \triangleq \langle f, g\rangle_{L^2(X,dx)} + \sum_{|\gamma|=l} \langle D^\gamma f, D^\gamma g\rangle_{L^2(X,dx)}.$$

Also note that the space $C_b^l(X)$ is a Banach space with respect to the norm

$$\|f\|_{C_b^l(X)} \triangleq \sup_{x\in X} |f(x)| + \sum_{|\gamma|=l} \sup_{x\in X} |D^\gamma f(x)|.$$

As explained in [15], since X is bounded, we have $C_b^l(X) \subset \mathcal{H}^l$ and $\|f\|_{\mathcal{H}^l} \leq C(l)\|f\|_{C_b^l(X)}$. Via Corollary 21 of section 4.6 from [16], if $m \in \mathbb{N}$ and $l - m > d/2$, then we also have:

$$\mathcal{H}^l \subset C_b^m(X) \quad \text{and} \quad \|f\|_{C_b^m(X)} \leq C(l, m)\, \|f\|_{\mathcal{H}^l}. \tag{B.1}$$

Following [15], if one takes $s \triangleq \lfloor d/2 \rfloor + 1$, then using (B.1) with $l = s$ and $m = 0$ we see that $\mathcal{H}^s$ is a RKHS with a continuous, real valued, bounded kernel h_s.

Appendix B.2. Additional operators

In this section we define several operators that will bridge the gap between the matrix $\mathbb{A}_\alpha$ and the operator A_α. All of these definitions are based on those from [15] for the asymmetrical diffusion operators (i.e. P). To start, define the empirical density maps $m_{\alpha,n} : X \to \mathbb{R}$ in terms of the samples $X_n = \{x^{(1)}, \ldots, x^{(n)}\}$ as

$$m_{\alpha,n}(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} k_\alpha(x, x^{(i)}), \quad \text{for all } \alpha \in \mathcal{I}, \;\; x \in X.$$

Note that $m_{\alpha,n}(x^{(i)}) = \mathbb{D}_\alpha[i, i]$. We also define the empirical kernels $a_{\alpha,n} : X \times X \to \mathbb{R}$ as

$$a_{\alpha,n}(x, y) \triangleq \frac{k_\alpha(x, y)}{\sqrt{m_{\alpha,n}(x)}\sqrt{m_{\alpha,n}(y)}}, \quad \text{for all } \alpha \in \mathcal{I}, \;\; x, y \in X.$$

We then have the following lemma from [15], adapted for symmetric diffusion operators.

Lemma Appendix B.1 (Lemma 16 from [15])**.** *Assume that (X, μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the conditions of Assumption 2. Then, for all $\alpha \in \mathcal{I}$ and for all $x \in X$,*

$$k_\alpha(x, \cdot), m_\alpha, m_{\alpha,n}, \frac{1}{m_\alpha}, \frac{1}{m_{\alpha,n}} \in C_b^{d+1}(X) \subset \mathcal{H}^{d+1} \subset \mathcal{H}^s,$$

$$\|k_\alpha(x, \cdot)\|_{C_b^{d+1}(X)}, \|m_\alpha\|_{C_b^{d+1}(X)}, \left\|m_{\alpha,n}\right\|_{C_b^{d+1}(X)}, \left\|\frac{1}{m_\alpha}\right\|_{C_b^{d+1}(X)}, \left\|\frac{1}{m_{\alpha,n}}\right\|_{C_b^{d+1}(X)} \leq C(\alpha, d),$$

$$a_\alpha(x, \cdot), a_{\alpha,n}(x, \cdot) \in C_b^{d+1}(X) \subset \mathcal{H}^{d+1} \subset \mathcal{H}^s,$$

$$\|a_\alpha(x, \cdot)\|_{\mathcal{H}^s}, \left\|a_{\alpha,n}(x, \cdot)\right\|_{\mathcal{H}^s} \leq C(\alpha, d).$$

Lemma Appendix B.1 allows one to define the operators $A_{\alpha,\mathcal{H}^s} : \mathcal{H}^s \to \mathcal{H}^s$ and $A_{\alpha,n} : \mathcal{H}^s \to \mathcal{H}^s$,

$$(A_{\alpha,\mathcal{H}^s} f)(x) \triangleq \int_X a_\alpha(x,y)\,\langle f, h_s(y,\cdot)\rangle_{\mathcal{H}^s}\, d\mu(y), \qquad \text{for all } \alpha \in \mathcal{I},\ \ f \in \mathcal{H}^s,$$

$$(A_{\alpha,n} f)(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} a_{\alpha,n}(x,x^{(i)})\,\langle f, h_s(x^{(i)},\cdot)\rangle_{\mathcal{H}^s}, \qquad \text{for all } \alpha \in \mathcal{I},\ \ f \in \mathcal{H}^s.$$

We also define similar operators $T_{\mathcal{H}^s} : \mathcal{H}^s \to \mathcal{H}^s$ and $T_n : \mathcal{H}^s \to \mathcal{H}^s$, but in terms of the reproducing kernel h_s.

$$(T_{\mathcal{H}^s} f)(x) \triangleq \int_X h_s(x,y)\,\langle f, h_s(y,\cdot)\rangle_{\mathcal{H}^s}\, d\mu(y), \qquad \text{for all } f \in \mathcal{H}^s,$$

$$(T_n f)(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} h_s(x,x^{(i)})\,\langle f, h_s(x^{(i)},\cdot)\rangle_{\mathcal{H}^s}, \qquad \text{for all } f \in \mathcal{H}^s.$$

The above operators, as well as A_α and $\mathbb{A}_\alpha$, can be decomposed in terms of the appropriate restriction and extension operators. We begin with the two restriction operators, $R_{\mathcal{H}^s} : \mathcal{H}^s \to L^2(X,\mu)$ and $R_n : \mathcal{H}^s \to \mathbb{R}^n$.

$$(R_{\mathcal{H}^s} f)(x) \triangleq \langle f, h_s(x,\cdot)\rangle_{\mathcal{H}^s}, \qquad \text{for } \mu \text{ a.e. } x \in X, \text{ for all } f \in \mathcal{H}^s,$$
$$R_n f \triangleq (f(x^{(1)}),\ldots,f(x^{(n)})), \qquad \text{for all } f \in \mathcal{H}^s.$$

For each $\alpha \in \mathcal{I}$ we also have two extension operators, $E_{\alpha,\mathcal{H}^s} : L^2(X,\mu) \to \mathcal{H}^s$ and $E_{\alpha,n} : \mathbb{R}^n \to \mathcal{H}^s$, where

$$(E_{\alpha,\mathcal{H}^s} f)(x) \triangleq \int_X a_\alpha(x,y) f(y)\, d\mu(y), \qquad \text{for all } x \in X,\ \ f \in L^2(X,\mu),$$

$$(E_{\alpha,n} v)(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} v[i]\, a_{\alpha,n}(x,x^{(i)}), \qquad \text{for all } x \in X,\ \ v \in \mathbb{R}^n.$$

Using these operators, one can easily show the following identities:

$$\begin{aligned} A_\alpha &= R_{\mathcal{H}^s} E_{\alpha,\mathcal{H}^s} \quad \text{and} \quad A_{\alpha,\mathcal{H}^s} = E_{\alpha,\mathcal{H}^s} R_{\mathcal{H}^s}, \\ \mathbb{A}_\alpha &= R_n E_{\alpha,n} \quad \text{and} \quad A_{\alpha,n} = E_{\alpha,n} R_n, \\ T_{\mathcal{H}^s} &= R^*_{\mathcal{H}^s} R_{\mathcal{H}^s} \quad \text{and} \quad T_n = R_n^* R_n. \end{aligned} \tag{B.2}$$

Appendix B.3. Similarity between empirical and continuous operators

Here we collect remaining results that we shall need that involve the similarity between the empirical and continuous versions of the previously defined operators and functions. All of these results can be found in [14, 15].

Theorem Appendix B.2 ([14], also Theorem 7 from [15])**.** *Suppose that* (X,μ) *and* $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ *satisfy the conditions of Assumption 2. Let* $n \in \mathbb{N}$ *and sample* $X_n = \{x^{(1)},\ldots,x^{(n)}\} \subset X$ *i.i.d.*

according to μ; also let $\tau > 0$. Then the operators $T_{\mathcal{H}^s}$ and T_n are Hilbert-Schmidt, and with probability $1 - 2e^{-\tau}$,

$$\|T_{\mathcal{H}^s} - T_n\|_{HS} \leq C(d)\frac{\sqrt{\tau}}{\sqrt{n}}.$$

Theorem Appendix B.3 (Theorem 15 from [15])**.** *Suppose that (X, μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the conditions of Assumption 2. Let $n \in \mathbb{N}$ and sample $X_n = \{x^{(1)}, \ldots, x^{(n)}\} \subset X$ i.i.d. according to μ; also let $\tau > 0$ and $\alpha \in \mathcal{I}$. Then the operators $A_{\alpha,\mathcal{H}^s}$ and $A_{\alpha,n}$ are Hilbert-Schmidt, and with probability $1 - 2e^{-\tau}$,*

$$\left\|A_{\alpha,\mathcal{H}^s} - A_{\alpha,n}\right\|_{HS} \leq C(\alpha, d)\frac{\sqrt{\tau}}{\sqrt{n}}.$$

Lemma Appendix B.4 (Lemma 18 from [15])**.** *Suppose that (X, μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the conditions of Assumption 2. Let $n \in \mathbb{N}$ and sample $X_n = \{x^{(1)}, \ldots, x^{(n)}\} \subset X$ i.i.d. according to μ; also let $\tau > 0$ and $\alpha \in \mathcal{I}$. Then, with probability $1 - 2e^{-\tau}$,*

$$\left\|m_\alpha - m_{\alpha,n}\right\|_{\mathcal{H}^{d+1}} \leq C(\alpha, d)\frac{\sqrt{\tau}}{\sqrt{n}}.$$

Appendix B.4. Proof of Theorem 5.1

In this section we prove Theorem 5.1, which we restate here.

Theorem Appendix B.5 (Theorem 5.1)**.** *Suppose that (X, μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the conditions of Assumption 2. Let $n \in \mathbb{N}$ and sample $X_n = \{x^{(1)}, \ldots, x^{(n)}\} \subset X$ i.i.d. according to μ; also let $t \in \mathbb{N}$, $\tau > 0$, and $\alpha, \beta \in \mathcal{I}$. Then, with probability $1 - 2e^{-\tau}$,*

$$\left|D^{(t)}(x_\alpha^{(i)}, x_\beta^{(j)}) - D_n^{(t)}(x_\alpha^{(i)}, x_\beta^{(j)})\right| \leq C(\alpha, \beta, d, t)\frac{\sqrt{\tau}}{\sqrt{n}}, \quad \textit{for all } i, j = 1, \ldots, n.$$

Proof of Theorem 5.1. First an additional piece of notation. Recall the d-dimensional index $\gamma = (\gamma_1, \ldots, \gamma_d)$. Let $\partial_x^\gamma a_\alpha$ denote the γ^{th} partial derivative of a_α with respect to the variable x.

We begin with the empirical diffusion distance. Recall that $D_n^{(t)}(x_\alpha^{(i)}, x_\beta^{(j)})^2 = n^2\|\mathbb{A}_\alpha^t[i, \cdot] - \mathbb{A}_\beta^t[j, \cdot]\|_{\mathbb{R}^n}^2$. For each $i = 1, \ldots, n$, define the vector $e^{(i)} \in \mathbb{R}^n$ as

$$e^{(i)}[j] \triangleq \begin{cases} 1, & \text{if } j = i, \\ 0, & \text{if } j \neq i, \end{cases} \qquad \text{for all } j = 1, \ldots, n.$$

We then have

$$\begin{aligned} D_n^{(t)}(x_\alpha^{(i)}, y_\beta^{(j)})^2 &= n^2\left\|\mathbb{A}_\alpha^t e^{(i)} - \mathbb{A}_\beta^t e^{(j)}\right\|_{\mathbb{R}^n}^2 \\ &= n^2\langle\mathbb{A}_\alpha^t e^{(i)}, \mathbb{A}_\alpha^t e^{(i)}\rangle_{\mathbb{R}^n} + n^2\langle\mathbb{A}_\beta^t e^{(j)}, \mathbb{A}_\beta^t e^{(j)}\rangle_{\mathbb{R}^n} - 2n^2\langle\mathbb{A}_\alpha^t e^{(i)}, \mathbb{A}_\beta^t e^{(j)}\rangle_{\mathbb{R}^n}. \end{aligned} \tag{B.3}$$

A similar expression can be had for the continuous diffusion distance. By assumption 2, $k_\alpha \in C_b^{d+1}(X \times X)$ and $k_\alpha \geq C_1(\alpha)$. These imply that $a_\alpha \in C_b^{d+1}(X \times X)$. We can then apply Mercer's Theorem to get that

$$a_\alpha^{(t)}(x, y) = \sum_{\ell\geq 1}\left(\lambda_\alpha^{(\ell)}\right)^t \psi_\alpha^{(\ell)}(x)\,\psi_\alpha^{(\ell)}(y), \quad \text{for all } (x, y) \in X \times X, \tag{B.4}$$

with absolute convergence and uniform convergence on compact subsets of X. In fact, since A_α is also trace class, we can get uniform convergence on all of X. Indeed,

$$\mathrm{Tr}(A_\alpha) = \sum_{\ell \geq 1} \lambda_\alpha^{(\ell)} < \infty.$$

Therefore, for all $\varepsilon > 0$ and for each $\alpha \in \mathcal{I}$, there exists $N(\varepsilon, \alpha) \in \mathbb{N}$ such that

$$\sum_{\ell > N(\varepsilon,\alpha)} \left(\lambda_\alpha^{(\ell)}\right)^t < \varepsilon.$$

Furthermore, since a_α is bounded, $\psi_\alpha^{(\ell)} \leq C_2(\alpha)$ for all $\ell \geq 1$. Therefore,

$$\sum_{\ell > N(\varepsilon,\alpha)} \left(\lambda_\alpha^{(\ell)}\right)^t \psi_\alpha^{(\ell)}(x)\, \psi_\alpha^{(\ell)}(y) \leq C_2(\alpha) \sum_{\ell > N(\varepsilon,\alpha)} \left(\lambda_\alpha^{(\ell)}\right)^t < C_2(\alpha)\, \varepsilon, \quad \text{for all } (x,y) \in X \times X. \tag{B.5}$$

Now define a family of functions $\varphi_\alpha^{(N,i)} \in L^2(X,\mu)$ for all $N \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$,

$$\varphi_\alpha^{(N,i)}(x) \triangleq \sum_{\ell=1}^{N} \psi_\alpha^{(\ell)}(x^{(i)})\, \psi_\alpha^{(\ell)}(x).$$

We claim that

$$\left| a_\alpha^{(t)}(x^{(i)}, x) - A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)}(x) \right| < C_2(\alpha)\, \varepsilon, \quad \text{for all } x \in X. \tag{B.6}$$

Indeed,

$$\begin{aligned}
A_\alpha^t \varphi_\alpha^{(N,i)}(x) &= \int_X a_\alpha^{(t)}(x,y)\, \varphi_\alpha^{(N,i)}(y)\, d\mu(y), \\
&= \int_X \left(\sum_{m \geq 1} \left(\lambda_\alpha^{(m)}\right)^t \psi_\alpha^{(m)}(x)\, \psi_\alpha^{(m)}(y) \right) \left(\sum_{\ell=1}^{N} \psi_\alpha^{(\ell)}(x^{(i)})\, \psi_\alpha^{(\ell)}(y) \right) d\mu(y), \\
&= \sum_{m \geq 1} \sum_{\ell=1}^{N} \left(\lambda_\alpha^{(m)}\right)^t \psi_\alpha^{(m)}(x)\, \psi_\alpha^{(\ell)}(x^{(i)}) \int_X \psi_\alpha^{(m)}(y)\, \psi_\alpha^{(\ell)}(y)\, d\mu(y), \\
&= \sum_{\ell=1}^{N} \left(\lambda_\alpha^{(\ell)}\right)^t \psi_\alpha^{(\ell)}(x^{(i)})\, \psi_\alpha^{(\ell)}(x).
\end{aligned} \tag{B.7}$$

Therefore, using (B.4), (B.7), and (B.5), we obtain

$$\left| a_\alpha^{(t)}(x^{(i)}, x) - A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)}(x) \right| = \left| \sum_{\ell > N(\varepsilon,\alpha)} \left(\lambda_\alpha^{(\ell)}\right)^t \psi_\alpha^{(\ell)}(x^{(i)})\, \psi_\alpha^{(\ell)}(x) \right| < C_2(\alpha)\, \varepsilon,$$

and so (B.6) holds.

Using (B.6), it not hard to see that

$$\left| D^{(t)}(x_\alpha^{(i)}, y_\beta^{(j)}) - \left\| A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)} - A_\beta^t \varphi_\beta^{(N\varepsilon,\beta),i)} \right\|_{L^2(X,\mu)} \right| \leq C_3(\alpha,\beta)\, \varepsilon.$$

Thus it is enough to consider $\|A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)} - A_\beta^t \varphi_\beta^{(N(\varepsilon,\beta),j)}\|_{L^2(X,\mu)}$. Expanding the square of this quantity one has

$$\left\|A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)} - A_\beta^t \varphi_\beta^{(N(\varepsilon,\beta),j)}\right\|_{L^2(X,\mu)}^2 = \langle A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)}, A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)}\rangle_{L^2(X,\mu)} + \langle A_\beta^t \varphi_\beta^{(N(\varepsilon,\beta),j)}, A_\beta^t \varphi_\beta^{(N(\varepsilon,\beta),j)}\rangle_{L^2(X,\mu)} - 2\langle A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)}, A_\beta^t \varphi_\beta^{(N(\varepsilon,\beta),j)}\rangle_{L^2(X,\mu)}. \tag{B.8}$$

The three inner products in (B.3) correspond to the three inner products in (B.8). We aim to show that each pair is nearly identical. We will do so explicitly for the pair $n^2\langle \mathbb{A}_\alpha^t e^{(i)}, \mathbb{A}_\beta^t e^{(j)}\rangle_{\mathbb{R}^n}$ and $\langle A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)}, A_\beta^t \varphi_\beta^{(N(\varepsilon,\beta),j)}\rangle_{L^2(X,\mu)}$; the other two pairs are simply special cases of this one. We begin with the discrete inner product, for which we have the following with probability $1 - 2e^{-\tau}$:

$$\begin{aligned} n^2\langle \mathbb{A}_\alpha^t e^{(i)}, \mathbb{A}_\beta^t e^{(j)}\rangle_{\mathbb{R}^n} &= n^2\langle (R_n E_{\alpha,n})^t e^{(i)}, (R_n E_{\beta,n})^t e^{(j)}\rangle_{\mathbb{R}^n} && \text{(B.9)}\\ &= n^2\langle (E_{\alpha,n} R_n)^{t-1} E_{\alpha,n} e^{(i)}, R_n^* R_n (E_{\beta,n} R_n)^{t-1} E_{\beta,n} e^{(j)}\rangle_{\mathcal{H}^s}\\ &= \langle A_{\alpha,n}^{t-1} a_{\alpha,n}(x^{(i)},\cdot), T_n A_{\beta,n}^{t-1} a_{\beta,n}(x^{(j)},\cdot)\rangle_{\mathcal{H}^s} && \text{(B.10)}\\ &\le \langle A_{\alpha,\mathcal{H}^s}^{t-1} a_{\alpha,n}(x^{(i)},\cdot), T_{\mathcal{H}^s} A_{\beta,\mathcal{H}^s}^{t-1} a_{\beta,n}(x^{(j)},\cdot)\rangle_{\mathcal{H}^s} + C(\alpha,\beta,d,t)\frac{\sqrt{\tau}}{\sqrt{n}}, && \text{(B.11)} \end{aligned}$$

where (B.9) follows from (B.2), (B.10) follows from (B.2) and the definitions of $E_{\alpha,n}$ and $e^{(i)}$, and (B.11) follows from Lemma Appendix B.1, Theorem Appendix B.2, Theorem Appendix B.3, and the Cauchy-Schwarz inequality. Since the argument is symmetric, we have, with probability $1 - 2e^{-\tau}$,

$$\left|n^2\langle \mathbb{A}_\alpha^t e^{(i)}, \mathbb{A}_\beta^t e^{(j)}\rangle_{\mathbb{R}^n} - \langle A_{\alpha,\mathcal{H}^s}^{t-1} a_{\alpha,n}(x^{(i)},\cdot), T_{\mathcal{H}^s} A_{\beta,\mathcal{H}^s}^{t-1} a_{\beta,n}(x^{(j)},\cdot)\rangle_{\mathcal{H}^s}\right| \le C(\alpha,\beta,d,t)\frac{\sqrt{\tau}}{\sqrt{n}}. \tag{B.12}$$

Now return to the continuous inner product. With probability $1 - 2e^{-\tau}$, we have:

$$\begin{aligned} \langle A_\alpha^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)}, A_\beta^t \varphi_\beta^{(N(\varepsilon,\beta),j)}\rangle_{L^2(X,\mu)} &= \langle (R_{\mathcal{H}^s} E_{\alpha,\mathcal{H}^s})^t \varphi_\alpha^{(N(\varepsilon,\alpha),i)}, (R_{\mathcal{H}^s} E_{\beta,\mathcal{H}^s})^t \varphi_\beta^{(N(\varepsilon,\beta),j)}\rangle_{L^2(X,\mu)} && \text{(B.13)}\\ &= \langle (E_{\alpha,\mathcal{H}^s} R_{\mathcal{H}^s})^{t-1} E_{\alpha,\mathcal{H}^s} \varphi_\alpha^{(N(\varepsilon,\alpha),i)}, R_{\mathcal{H}^s}^* R_{\mathcal{H}^s} (E_{\beta,\mathcal{H}^s} R_{\mathcal{H}^s})^{t-1} E_{\beta,\mathcal{H}^s} \varphi_\beta^{(N(\varepsilon,\beta),j)}\rangle_{\mathcal{H}^s}\\ &= \langle A_{\alpha,\mathcal{H}^s}^{t-1} E_{\alpha,\mathcal{H}^s} \varphi_\alpha^{(N(\varepsilon,\alpha),i)}, T_{\mathcal{H}^s} A_{\beta,\mathcal{H}^s}^{t-1} E_{\beta,\mathcal{H}^s} \varphi_\beta^{(N(\varepsilon,\beta),j)}\rangle_{\mathcal{H}^s}, && \text{(B.14)} \end{aligned}$$

where (B.13) and (B.14) both follow from (B.2).

Examining (B.12) and (B.14), it is clear that to complete the proof we must bound the quantity $\|a_{\alpha,n}(x^{(i)},\cdot) - E_{\alpha,\mathcal{H}^s}\varphi_\alpha^{(N(\varepsilon,\alpha),i)}\|_{\mathcal{H}^s}$. We break it into two parts:

$$\left\|a_{\alpha,n}(x^{(i)},\cdot) - E_{\alpha,\mathcal{H}^s}\varphi_\alpha^{(N(\varepsilon,\alpha),i)}\right\|_{\mathcal{H}^s} \le \left\|a_{\alpha,n}(x^{(i)},\cdot) - a_\alpha(x^{(i)},\cdot)\right\|_{\mathcal{H}^s} + \left\|a_\alpha(x^{(i)},\cdot) - E_{\alpha,\mathcal{H}^s}\varphi_\alpha^{(N(\varepsilon,\alpha),i)}\right\|_{\mathcal{H}^s}. \tag{B.15}$$

For the first part, some simple manipulations give:

$$a_{\alpha,n}(x^{(i)},x) - a_\alpha(x^{(i)},x) = f_{\alpha,n}^{(i)}(x) + g_{\alpha,n}^{(i)}(x),$$

where

$$f_{\alpha,n}^{(i)}(x) = \frac{k_\alpha(x^{(i)},x)\,(\sqrt{m_\alpha(x)} - \sqrt{m_{\alpha,n}(x)})}{\sqrt{m_{\alpha,n}(x^{(i)})}\sqrt{m_{\alpha,n}(x)}\sqrt{m_\alpha(x)}}$$

and

$$g^{(i)}_{\alpha,n}(x) = \frac{k_\alpha(x^{(i)},x)\,(\sqrt{m_{\alpha,n}(x^{(i)})} - \sqrt{m_\alpha(x^{(i)})})}{\sqrt{m_{\alpha,n}(x^{(i)})}\,\sqrt{m_\alpha(x^{(i)})}\,\sqrt{m_\alpha(x)}}.$$

For the first of these two functions, using Lemma Appendix B.1 and Lemma Appendix B.4 it is easy to see that $\|f^{(i)}_{\alpha,n}\|_{\mathcal{H}^s} \le C(\alpha,d)\frac{\sqrt{\tau}}{\sqrt{n}}$ with probability $1-2e^{-\tau}$. For $g^{(i)}_{\alpha,n}$, note that

$$\begin{aligned}\left|m_{\alpha,n}(x^{(i)}) - m_\alpha(x^{(i)})\right| &\le \sup_{x\in X}\left|m_{\alpha,n}(x) - m_\alpha(x)\right| \\ &= \left\|m_{\alpha,n} - m_\alpha\right\|_{C^0_b(X)} \\ &\le C(d)\left\|m_{\alpha,n} - m_\alpha\right\|_{\mathcal{H}^{d+1}} \\ &\le C(\alpha,d)\frac{\sqrt{\tau}}{\sqrt{n}}, \end{aligned} \tag{B.16}$$

where in (B.16) we once again used Lemma Appendix B.4. Thus $\|g^{(i)}_{\alpha,n}\|_{\mathcal{H}^s} \le C(\alpha,d)\frac{\sqrt{\tau}}{\sqrt{n}}$ with probability $1-2e^{-\tau}$, and so we have bounded the first term on the right hand side of (B.15). For the second term on the right hand side of (B.15), recall the definition of $\|\cdot\|_{\mathcal{H}^s}$. If we can bound $\|\partial^\gamma_x a_\alpha(x^{(i)},\cdot) - \partial^\gamma_x E_{\alpha,\mathcal{H}^s}\varphi^{(N(\varepsilon,\alpha),i)}_\alpha\|_{L^2(X,dx)}$, where $\gamma = 0$ (i.e., no derivative) or $|\gamma| = s$, then we will have bounded this term as well. Note that $a_\alpha \in C^{d+1}_b(X\times X)$ implies that $\psi^{(\ell)}_\alpha \in C^s_b(X)$ for all $\ell \ge 1$. Furthermore, the derivative $\partial^\gamma_x a_\alpha(x^{(i)},\cdot)$ can be computed term by term from (B.4). Thus, using nearly the same argument we used to show (B.6), one can show that

$$\left|\partial^\gamma_x a_\alpha(x^{(i)},x) - \partial^\gamma_x E_{\alpha,\mathcal{H}^s}\varphi^{(N(\varepsilon,\alpha),i)}_\alpha(x)\right| < C_4(\alpha)\,\varepsilon, \quad \text{for all } x\in X,\ |\gamma|\le s. \tag{B.17}$$

Using (B.17), we have:

$$\left\|\partial^\gamma_x a_\alpha(x^{(i)},\cdot) - \partial^\gamma_x E_{\alpha,\mathcal{H}^s}\varphi^{(N(\varepsilon,\alpha),i)}_\alpha\right\|_{L^2(X,dx)} \le \sqrt{|X|}\,C_4(\alpha)\,\varepsilon,$$

where $|X|$ denotes the Lebesgue measure of X. Since X was assumed to be bounded, we have $|X| \le C$. Returning to (B.15), we have now shown that:

$$\left\|a_{\alpha,n}(x^{(i)},\cdot) - E_{\alpha,\mathcal{H}^s}\varphi^{(N(\varepsilon,\alpha),i)}_\alpha\right\|_{\mathcal{H}^s} \le C(\alpha,d)\frac{\sqrt{\tau}}{\sqrt{n}} + C\varepsilon.$$

Taking $\varepsilon = \frac{\sqrt{\tau}}{\sqrt{n}}$ completes the proof. □

Appendix B.5. Proof of Theorem 5.2

Finally, we prove Theorem 5.2.

Theorem Appendix B.6 (Theorem 5.2)**.** *Suppose that (X,μ) and $\{k_\alpha\}_{\alpha\in\mathcal{I}}$ satisfy the conditions of Assumption 2. Let $n\in\mathbb{N}$ and sample $X_n = \{x^{(1)},\dots,x^{(n)}\} \subset X$ i.i.d. according to μ; also let $t\in\mathbb{N}$, $\tau > 0$, and $\alpha,\beta\in\mathcal{I}$. Then, with probability $1-2e^{-\tau}$,*

$$\left|\mathcal{D}^{(t)}(\Gamma_\alpha,\Gamma_\beta) - \mathcal{D}^{(t)}_n(\Gamma_{\alpha,n},\Gamma_{\beta,n})\right| \le C(\alpha,\beta,d,t)\frac{\sqrt{\tau}}{\sqrt{n}}.$$

Proof. Recall that $\mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta) = \|A^t_\alpha - A^t_\beta\|_{HS}$. From Proposition 13 in [15], we know that $\lambda \in (0, 1]$ is an eigenvalue of A_α if and only if it is an eigenvalue of $A_{\alpha,\mathcal{H}^s}$. Using the same ideas, one can show that $\lambda' \neq 0$ is an eigenvalue of $A^t_\alpha - A^t_\beta$ if and only if it is an eigenvalue of $A^t_{\alpha,\mathcal{H}^s} - A^t_{\beta,\mathcal{H}^s}$. Therefore,

$$\left\|A^t_\alpha - A^t_\beta\right\|_{HS} = \left\|A^t_{\alpha,\mathcal{H}^s} - A^t_{\beta,\mathcal{H}^s}\right\|_{HS}.$$

Similarly, one can show that

$$\left\|\mathbb{A}^t_\alpha - \mathbb{A}^t_\beta\right\|_{HS} = \left\|A^t_{\alpha,n} - A^t_{\beta,n}\right\|_{HS}.$$

Thus, using the above and Theorem Appendix B.3 we have, with probability $1 - 2e^{-\tau}$,

$$\begin{aligned}
\mathcal{D}^{(t)}(\Gamma_\alpha, \Gamma_\beta) &= \|A^t_{\alpha,\mathcal{H}^s} - A^t_{\beta,\mathcal{H}^s}\|_{HS} \\
&\leq \|A^t_{\alpha,n} - A^t_{\beta,n}\|_{HS} + \|A^t_{\alpha,\mathcal{H}^s} - A^t_{\alpha,n}\|_{HS} + \|A^t_{\beta,\mathcal{H}^s} - A^t_{\beta,n}\|_{HS} \\
&\leq \mathcal{D}^{(t)}_n(\Gamma_{\alpha,n}, \Gamma_{\beta,n}) + C(\alpha, \beta, d, t)\frac{\sqrt{\tau}}{\sqrt{n}}.
\end{aligned}$$

Since the argument is symmetric, we get the desired inequality. □

References

[1] M. T. Eismann, J. Meola, R. C. Hardie, Hyperspectral change detection in the presence of diurnal and seasonal variations, IEEE Transactions on Geoscience and Remote Sensing 46 (2008) 237–249.

[2] R. R. Coifman, S. Lafon, Diffusion maps, Applied and Computational Harmonic Analysis 21 (2006) 5–30.

[3] U. Vaidya, G. Hagen, S. Lafon, A. Banaszuk, I. Mezić, R. R. Coifman, Comparison of systems using diffusion maps, in: Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference 2005, Seville, Spain, pp. 7931–7936.

[4] J. D. Lee, M. Maggioni, Multiscale analysis of time series of graphs, in: Proceedings of The 9th International Conference on Sampling Theory and Applications, Singapore.

[5] B. Gérard, Processus de Diffusion sur un Flot de Variétés Riemanniennes, Ph.D. thesis, L'Universite de Grenoble, 2010.

[6] F. Mémoli, A spectral notion of Gromov-Wasserstein distance and related methods, Applied and Computational Harmonic Analysis 30 (2011) 363–401.

[7] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[8] J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.

[9] D. L. Donoho, C. Grimes, Hessian eigenmaps: new locally lienar embedding techniques for high-dimensional data, Proceedings of the National Academy of Sciences of the United States of America 100 (2003) 5591–5596.

[10] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (2003) 1373–1396.

[11] J. Mercer, Functions of positive and negative type and their connection with the theory of integral equations, Philosophical Transactions of the Royal Society of London, Series A 209 (1909) 415–446.

[12] H. Q. Minh, P. Niyogi, Y. Yao, Mercer's theorem, feature maps, and smoothing, in: Conference on Learning Theory, Lecture Notes in Computer Science, Springer, Pittsburgh, Pennsylvania, USA, 2006, pp. 154–168.

[13] G. B. Folland, Real Analysis - Modern Techniques and their Applications, John Wiley & Sons, Inc, 2nd edition, 1999.

[14] E. D. Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, F. Odone, Learning from examples as an inverse problem, Journal of Machine Learning Research 6 (2005) 883–904.

[15] L. Rosasco, M. Belkin, E. D. Vito, On learning with integral operators, Journal of Machine Learning Research 11 (2010) 905–934.

[16] V. I. Burenkov, Sobolev Spaces on Domains, Teubner-Texte zur Mathematik, B.G. Teubner, Stuttgart-Leipzig, 1998.
[17] Z. Levnajić, I. Mezić, Ergodic theory and visualization I: Mesochronic plots for visualization of ergodic partition and invariant sets, Chaos 20 (2010) 033114.
[18] Z. Levnajić, I. Mezić, Ergodic theory and visualization II: Visualization of resonances and periodic sets, 2008. ArXiv:0808.2182v1.
[19] R. R. Coifman, M. Hirn, R. Lederman, Diffusion embeddings of parameterized difference equations, 2012. In preparation.
[20] A. Singer, H. tieng Wu, Vector diffusion maps and the connection Laplacian, Communications on Pure and Applied Mathematics 65 (2012) 1067–1144.
[21] G. Wolf, A. Averbuch, Linear-projection diffusion on smooth Euclidean submanifolds, Applied and Computational Harmonic Analysis (2012). In press.